PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-132289
(43)Date of publication of application : 09.05.2002

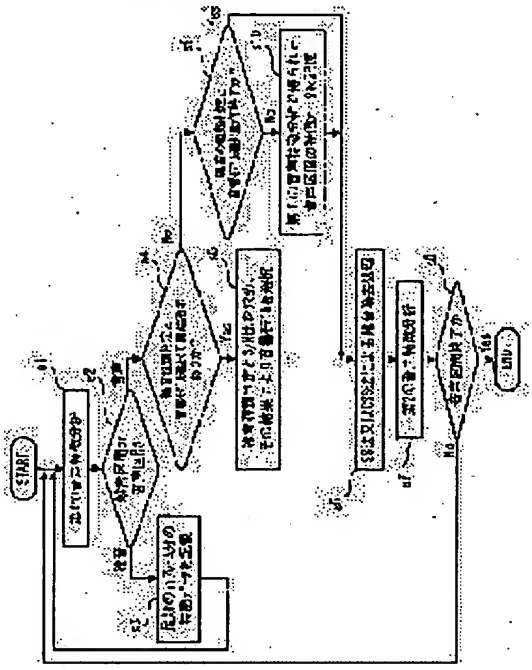
(51)Int.Cl. G10L 15/20
G10L 15/06
G10L 21/02

(21)Application number : 2000-322914
(22)Date of filing : 23.10.2000

(71)Applicant : SEIKO EPSON CORP
(72)Inventor : MIYAZAWA YASUNAGA

(54) SPEECH RECOGNITION METHOD AND RECORDING MEDIUM RECORDED WITH SPEECH RECOGNITION PROCESSING PROGRAM AS WELL AS SPEECH RECOGNIZER

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain a high recognition rate even under the environment where plural kinds of noises exist.
SOLUTION: Respective pieces of the speech data superposed with the noises of different kinds are subjected to noise removal by using a noise removing method (called as an SS method) of spectral subtraction and the acoustic models corresponding to the kinds of the noise formed by the characteristic vectors obtained by characteristic analysis of respective pieces of the speech data after the noise removal are prepared. In recognition, the speech data to be recognized is subjected to the first speech characteristic analysis and noise sections/speech sections are decided and in the case of the noise sections, the characteristic data thereof is saved (steps s1 to s3). In the case of the speech sections, the noise kinds are decided by the saved characteristic data and the corresponding to acoustic model is selected by the results thereof (step s5). The speech data to be recognized is subjected to the noise removal by the SS method and the speech data after the noise removal is subjected to the second speech characteristic analysis for obtaining the characteristic spectra for speech recognition (steps processing and s6 to s7).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

CLAIMS

[Claim(s)]

[Claim 1] The noise from which a kind differs creates each voice data on which it was superimposed for every kind of noise. By performing a normal mode rejection using the predetermined normal-mode-rejection technique, and using the feature vector of each of the voice data by which the normal mode rejection was carried out to each voice data superimposed on these noises The sound model group corresponding to the kind of noise is created, and it is held. at the time of recognition While judging the kind of noise on which it is superimposed to the voice data for recognition superimposed on noise and choosing a predetermined sound model out of the sound model group corresponding to the kind of the aforementioned noise based on the judgment result The speech recognition method characterized by performing a normal mode rejection using the aforementioned predetermined normal-mode-rejection method, and performing speech recognition to the voice data for recognition superimposed on the aforementioned noise using the sound model by which selection was carried out [aforementioned] to the feature vector of the voice data by which the normal mode rejection was carried out.

[Claim 2] The aforementioned normal-mode-rejection technique is the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique. in that case the aforementioned sound model group A normal mode rejection is performed using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique to each voice data superimposed on the noise from which the aforementioned kind differs. A feature vector is obtained from each voice data by which the normal mode rejection was carried out, and it is created using the feature vector. at the time of recognition The 1st voice feature analysis for obtaining the feature data in a frequency domain is performed to the voice data for recognition superimposed on the aforementioned noise. When the noise section or the voice section is judged and it is judged with it being the noise section using the feature-analysis result When judged with the analyzed feature data being saved and it being the voice section The kind of noise on which it is superimposed is judged with the feature data by which preservation was carried out [aforementioned]. Based on the judgment result, a predetermined sound model is chosen out of the sound model group prepared for kind correspondence of the aforementioned noise. A normal mode rejection is performed to the voice data for recognition superimposed on the aforementioned noise using the normal-mode-rejection technique by the aforementioned SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique. The speech recognition method according to claim 1 characterized by what the 2nd voice feature analysis for obtaining the feature data required for speech recognition is performed to the voice data by which the normal mode rejection was carried out, and speech recognition is performed for to the feature-analysis result using the sound model by which selection was carried out [aforementioned].

[Claim 3] The aforementioned normal-mode-rejection technique is the normal-mode-rejection technique by the cepstrum average normalizing method. in that case the aforementioned sound model A normal mode rejection is performed using the normal-mode-rejection technique by the cepstrum average normalizing method to each voice data superimposed on the noise from which the aforementioned kind differs. It is created using the feature vector of

opposed to the feature vector of the voice section by which preservation is carried out [aforementioned] to the voice section where it was superimposed on the aforementioned noise The speech recognition method characterized by what speech recognition is performed for using the sound model by which selection was carried out [aforementioned] to the feature vector which obtained the feature vector to the voice section concerned with the application of the cepstrum average normalizing method, and was obtained with the application of the cepstrum average normalizing method.

[Claim 7] In addition to the kind of noise, the sound model of noise correspondence of each above is used as the sound model also corresponding to the S/N ratio of two or more step story for every kind of each noise. the sound model in that case Each voice data on which each noise from which a S/N ratio differs for two or more kinds of every noises was made to superimpose for every kind of noise is created. A normal mode rejection is performed to this voice data using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique. As opposed to each of the voice data by which the normal mode rejection was carried out The speech recognition method according to claim 6 characterized by being the sound model of S/N ratio correspondence of two or more step story for every kind of each noise created using the feature vector of each voice data obtained with the application of the cepstrum average normalizing method.

[Claim 8] When the aforementioned sound model is a sound model also corresponding to the S/N ratio of two or more step story per kind of each noise, It adds to the processing which judges the kind of noise to the voice data for recognition superimposed on the aforementioned noise. The speech recognition method according to claim 7 characterized by performing processing which asks for a S/N ratio and choosing a sound model from the size of the noise of the noise section, and the size of the voice of the voice section based on the judged noise kind and the called-for S/N ratio.

[Claim 9] A certain specific kind from which a S/N ratio differs of noise creates each voice data on which it was superimposed for every S/N ratio. By performing a normal mode rejection using the predetermined normal-mode-rejection technique, and using the feature vector of each of the voice data by which the normal mode rejection was carried out to each voice data of these The sound model group corresponding to each S/N ratio is created, and it is held. at the time of recognition While judging the S/N ratio on which it is superimposed to the voice data for recognition superimposed on noise and choosing a predetermined sound model out of the sound model group corresponding to the aforementioned S/N ratio based on the judgment result The speech recognition method characterized by performing a normal mode rejection using the aforementioned predetermined normal-mode-rejection method, and performing speech recognition to the voice data for recognition superimposed on the aforementioned noise using the sound model by which selection was carried out [aforementioned] to the feature vector of the voice data by which the normal mode rejection was carried out.

[Claim 10] The aforementioned normal-mode-rejection technique is the speech recognition method according to claim 9 characterized by being the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique.

[Claim 11] The aforementioned normal-mode-rejection technique is the speech recognition method according to claim 9 characterized by being the normal-mode-rejection technique by the cepstrum average normalizing method.

[Claim 12] The noise from which a kind differs creates each voice data on which it was superimposed for every kind of noise. To each voice data superimposed on the noise from which these kinds differ, perform a normal mode rejection using the predetermined normal-mode-rejection technique, and each of the voice data by which the normal mode rejection was carried out by the feature vector obtained by carrying out feature-analysis processing The procedure of creating the sound model group corresponding to the kind of noise, and making a sound model group storage means memorizing it, The procedure which chooses a predetermined sound model out of the sound model group which judged the kind of noise on which it is superimposed to the voice data for recognition superimposed on noise, and was memorized by the aforementioned sound model group storage means based on the

for every kind of each noise. the sound model in that case Each noise from which a S/N ratio differs for two or more kinds of every noises creates each voice data on which it was superimposed for every kind of noise. A normal mode rejection is performed to voice data using the predetermined normal-mode-rejection technique. this each -- The record medium which recorded the speech recognition processing program of a publication on either of the claims 12-14 characterized by being a sound model corresponding to the S/N ratio of two or more step story for every kind of each noise created by the feature vector of each of the voice data by which the normal mode rejection was carried out.

[Claim 16] When the aforementioned sound model is a sound model also corresponding to the S/N ratio of two or more step story per kind of each noise, It adds to the processing which judges the kind of noise to the voice data for recognition superimposed on the aforementioned noise. The record medium which recorded the speech recognition processing program according to claim 15 characterized by performing processing which asks for a S/N ratio and choosing a sound model from the size of the noise of the noise section, and the size of the voice of the voice section based on the judged noise kind and the called-for S/N ratio.

[Claim 17] The noise from which a kind differs creates each voice data on which it was superimposed for every kind of noise. As opposed to each voice data superimposed on the noise from which these kinds differ A normal mode rejection is performed using the normal-mode-rejection technique by the SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method. With the application of the cepstrum average normalizing method, the feature vector to the voice section concerned is obtained to each of the voice data by which the normal mode rejection was carried out. by the feature vector The procedure of creating each sound model group dealing with noise, and making a sound model group storage means memorizing it, The procedure of performing the 1st voice feature analysis for obtaining the feature data in a frequency domain to the voice data for recognition superimposed on noise, When the noise section or the voice section is judged and it is judged with it being the noise section using the feature analysis result The procedure of saving the analyzed feature data, and when it is judged with it being the voice section The procedure of performing a normal mode rejection using the normal-mode-rejection technique by the describing [above] SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method to the voice section, The procedure of performing the 2nd voice feature analysis processing for asking for a cepstrum coefficient to the data of the voice section by which the normal mode rejection was carried out, and saving the feature vector of the voice section, By the feature analytical data of the noise section by which preservation was carried out [aforementioned] after the voice section end The procedure which chooses a predetermined sound model out of the sound model group which judged the kind of noise on which it is superimposed and was prepared for the aforementioned noise correspondence based on the judgment result, The procedure of obtaining the feature vector to the voice section concerned with the application of the cepstrum average normalizing method to the feature vector of the voice section by which preservation is carried out [aforementioned] to the voice section where it was superimposed on the aforementioned noise, The record medium which recorded the speech recognition processing program characterized by including the procedure of performing speech recognition using the sound model by which selection was carried out [aforementioned] in the processing program to the feature vector obtained with the application of the cepstrum average normalizing method.

[Claim 18] In addition to the kind of noise, the sound model of noise correspondence of each above is used as the sound model also corresponding to the S/N ratio of two or more step story for every kind of each noise. the sound model in that case Each voice data on which each noise from which a S/N ratio differs for two or more kinds of every noises was made to superimpose for every kind of noise is created. A normal mode rejection is performed to this voice data using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique. As opposed to each of the voice data by which the normal mode rejection was carried out The record medium which recorded the speech recognition processing program according to claim 17 characterized by being the sound model of S/N ratio correspondence of two or more step story for every kind of each noise

group which obtained the feature vector from each voice data by which the normal mode rejection was carried out, was created using the feature vector, and was created by this, to each voice data superimposed on the noise from which the aforementioned kind differs. The 1st voice feature analysis means which performs the 1st voice feature analysis for obtaining the feature analytical data in a frequency domain to the voice data for recognition superimposed on the aforementioned noise. The noise section / a voice section judging means to save the feature data of the noise section for the feature data-storage means when the noise section or the voice section is judged and it judges with it being the noise section using the feature analysis result. When judged with it being the voice section, with the feature data by which preservation was carried out [aforementioned] A noise kind judging means to judge the kind of noise on which it is superimposed, and a sound model-selection means to choose a predetermined sound model based on the judgment result out of the aforementioned sound model group prepared for kind correspondence of the aforementioned noise, A normal-mode-rejection means to perform a normal mode rejection to the voice data for recognition superimposed on the aforementioned noise using the normal-mode-rejection technique by the describing [above] SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method, The 2nd voice feature analysis means which performs the 2nd voice feature analysis for obtaining the feature data required for speech recognition to the voice data by which the normal mode rejection was carried out, and a speech recognition means to perform speech recognition to the feature analysis result using the sound model by which selection was carried out [aforementioned].

[Claim 25] The normal-mode-rejection technique which the aforementioned normal-mode-rejection means performs is the normal-mode-rejection technique by the cepstrum average normalizing method. in that case the aforementioned sound model A normal mode rejection is performed using the normal-mode-rejection technique by the cepstrum average normalizing method to each voice data superimposed on the noise from which the aforementioned kind differs. A sound model group storage means to memorize the sound model group which was created using the feature vector of each voice data obtained by it, and was created by this, A feature analysis means to perform the feature analysis for asking for the feature vector showing a cepstrum coefficient from the voice data for recognition superimposed on the aforementioned noise, When the noise section or the voice section is judged and it judges with it being the noise section using the feature analysis result When it judges with the feature vector of the noise section being saved for the feature data-storage means, and it being the voice section By the feature vector of the noise section saved for the noise section / a voice section judging means to save the feature vector of the voice section for the feature analytical-data storage means, and this feature data-storage means A noise kind judging means to judge the kind of noise on which it is superimposed, and a sound model-selection means to choose a predetermined sound model based on the judgment result out of the aforementioned sound model group prepared for kind correspondence of the aforementioned noise, A normal-mode-rejection means to perform normal-mode-rejection processing to the voice section where it was superimposed on the aforementioned noise using the normal-mode-rejection technique by the cepstrum average normalizing method using the feature vector of the voice section by which preservation is carried out [aforementioned], A speech recognition means to perform speech recognition to the feature vector obtained by the normal-mode-rejection processing using the sound model by which selection was carried out [aforementioned], and the voice recognition unit according to claim 23 characterized by having..

[Claim 26] In addition to the kind of noise, the sound model of noise correspondence of each above is used as the sound model also corresponding to the S/N ratio of two or more step story for every kind of each noise. the sound model in that case Each noise from which a S/N ratio differs for two or more kinds of every noises creates each voice data on which it was superimposed for every kind of noise. A normal mode rejection is performed to voice data using the predetermined normal-mode-rejection technique. this each -- A voice recognition unit given in either of the claims 23-25 characterized by being a sound model corresponding to the S/N ratio of two or more step story for every kind of each noise created by the feature

according to claim 29 which performs processing which asks for a S/N ratio from the size of the noise of the noise section, and the size of the voice of the voice section, and is characterized by the aforementioned sound model-selection section choosing a sound model based on the judged noise kind and the called-for S/N ratio.

[Claim 31] The voice recognition unit characterized by providing the following. The sound model group corresponding to the S/N ratio by which each voice data on which it was superimposed for every kind of noise was created, and the noise from which a S/N ratio differs performed the normal mode rejection using the predetermined normal-mode-rejection technique, and was created by the feature vector of each of the voice data by which the normal mode rejection was carried out to each voice data of these. A sound model group storage means to memorize this sound model group. A S/N ratio judging means to judge the S/N ratio of the noise on which it is superimposed to the voice data for recognition superimposed on noise. A sound model group-selection means to choose a predetermined sound model out of the sound model group corresponding to the aforementioned S/N ratio based on the judgment result, a normal-mode-rejection means to perform a normal mode rejection to the voice data for recognition superimposed on the aforementioned noise using the aforementioned predetermined normal-mode-rejection method, and a speech recognition means perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of voice data by which the normal mode rejection was carried out.

[Claim 32] The aforementioned normal-mode-rejection technique is a voice recognition unit according to claim 31 characterized by being the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique.

[Claim 33] The aforementioned normal-mode-rejection technique is a voice recognition unit according to claim 31 characterized by being the normal-mode-rejection technique by the cepstrum average normalizing method.

[Translation done.]

data using the normal-mode-rejection method by the SUPEKUTORARU subtraction, the sound model for speech recognition learned using the voice data from which the noise was removed is created.

[0009] Thus, if speech recognition is performed using the created sound model about a certain specific noise, although a good result will be comparatively obtained under the environment where such noise exists, the other noise may sometimes exist [enough] depending on a situation. The recognition rate in that case falls with a natural thing.

[0010] Moreover, a recognition performance changes also with sizes of the S/N ratio which is a ratio of a sound signal and a noise signal which should actually be recognized besides the kind of noise.

[0011] Then, this invention can obtain the high recognition performance corresponding to the kind of noise, or the size of a S/N ratio, and aims at moreover making realization possible by the cheap hardware using low CPU of arithmetic proficiency.

[0012]

[Means for Solving the Problem] In order to attain the purpose mentioned above, it is the noise from which a kind's differs creating each voice data on which it was superimposed for every kind of noise, and the speech recognition method of this invention performing a normal mode rejection to each voice data superimposed on these noises using the predetermined normal-mode-rejection technique, and using the feature vector of each of the voice data by which the normal mode rejection's was carried out, and the sound model group corresponding to the kind of noise is created, and it is held. And while judging the kind of noise on which it is superimposed to the voice data for recognition superimposed on noise and choosing a predetermined sound model out of the sound model group corresponding to the kind of the aforementioned noise based on the judgment result at the time of recognition To the voice data for recognition superimposed on the aforementioned noise, a normal mode rejection is performed using the aforementioned predetermined normal-mode-rejection method, and it is made to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of the voice data by which the normal mode rejection was carried out.

[0013] Moreover, the record medium which recorded the speech recognition processing program of this invention The noise from which a kind differs creates each voice data on which it was superimposed for every kind of noise. To each voice data superimposed on the noise from which these kinds differ, perform a normal mode rejection using the predetermined normal-mode-rejection technique, and each of the voice data by which the normal mode rejection was carried out by the feature vector obtained by carrying out the feature analysis processing The procedure of creating the sound model group corresponding to the kind of noise, and making a sound model group storage means memorizing it, The procedure which chooses a predetermined sound model out of the sound model group which judged the kind of noise on which it is superimposed to the voice data for recognition superimposed on noise, and was memorized by the aforementioned sound model group storage means based on the judgment result, The procedure of performing a normal mode rejection using the aforementioned predetermined normal-mode-rejection method, and the procedure of performing speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of voice data by which the normal mode rejection was carried out are included in the processing program to the voice data for recognition superimposed on the aforementioned noise.

[0014] In that case, to each voice data superimposed on the noise from which the aforementioned kind differs, the aforementioned normal-mode-rejection technique is the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique, and the aforementioned sound model group performs a normal mode rejection using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique, it obtains a feature vector from each voice data by which the normal mode rejection was carried out, and is created in each [these] invention using the feature vector. And at the time of recognition, the voice data for recognition superimposed on the aforementioned noise

application of the cepstrum average normalizing method, the feature vector to the voice section concerned is obtained to each of the voice data by which the normal mode rejection was carried out, by the feature vector, each sound model group dealing with noise is created, and it is saved. And at the time of recognition, the voice data for recognition superimposed on noise is received. When the 1st voice feature analysis for obtaining the feature data in a frequency domain is performed, the noise section or the voice section is judged using the feature-analysis result and it is judged with it being the noise section When judged with the analyzed feature data being saved and it being the voice section A normal mode rejection is performed using the normal-mode-rejection technique by the aforementioned SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique to the voice section. With the feature data of the noise section by which performed 2nd voice feature-analysis processing for asking for a cepstrum coefficient to the data of the voice section by which the normal mode rejection was carried out, and saved the feature vector of the voice section, and preservation was carried out [aforementioned] after the voice section end Judge the kind of noise on which it is superimposed and a predetermined sound model is chosen out of the sound model group prepared for the aforementioned noise correspondence based on the judgment result. As opposed to the feature vector of the voice section by which preservation is carried out [aforementioned] to the voice section where it was superimposed on the aforementioned noise It is made to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector which obtained the feature vector to the voice section concerned with the application of the cepstrum average normalizing method, and was obtained with the application of the cepstrum average normalizing method.

[0019] Moreover, the record medium which recorded another speech recognition processing program The noise from which a kind differs creates each voice data on which it was superimposed for every kind of noise. As opposed to each voice data superimposed on the noise from which these kinds differ A normal mode rejection is performed using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique. With the application of the cepstrum average normalizing method, the feature vector to the voice section concerned is obtained to each of the voice data by which the normal mode rejection was carried out. by the feature vector The procedure of creating each sound model group dealing with noise, and making a sound model group storage means memorizing it, The procedure of performing the 1st voice feature analysis for obtaining the feature data in a frequency domain to the voice data for recognition superimposed on noise, When the noise section or the voice section is judged and it is judged with it being the noise section using the feature-analysis result The procedure of saving the analyzed feature data, and when it is judged with it being the voice section The procedure of performing a normal mode rejection using the normal-mode-rejection technique by the aforementioned SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique to the voice section, The procedure of performing 2nd voice feature-analysis processing for asking for a cepstrum coefficient to the data of the voice section by which the normal mode rejection was carried out, and saving the feature vector of the voice section, With the feature-analysis data of the noise section by which preservation was carried out [aforementioned] after the voice section end The procedure which chooses a predetermined sound model out of the sound model group which judged the kind of noise on which it is superimposed and was prepared for the aforementioned noise correspondence based on the judgment result, The procedure of obtaining the feature vector to the voice section concerned with the application of the cepstrum average normalizing method to the feature vector of the voice section by which preservation is carried out [aforementioned] to the voice section where it was superimposed on the aforementioned noise, The procedure of performing speech recognition using the sound model by which selection was carried out [aforementioned] is included in the processing program to the feature vector obtained with the application of the cepstrum average normalizing method.

[0020] In the record medium which recorded these speech recognition method and the speech recognition processing program the sound model of noise correspondence of each above In

each of the voice data by which the normal mode rejection was carried out,-A sound model group storage means to memorize this sound model group, and a noise judging means to judge the kind of noise on which it is superimposed to the voice data for recognition superimposed on noise, A sound model group-selection means to choose a predetermined sound model out of the sound model group corresponding to the kind of the aforementioned noise based on the judgment result, It is considering as composition with a normal-mode-rejection means to perform a normal mode rejection using the aforementioned predetermined normal-mode-rejection method, and a speech recognition means to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of voice data by which the normal mode rejection was carried out, to the voice data for recognition superimposed on the aforementioned noise.

[0026] In this voice recognition unit, the normal-mode-rejection technique which the aforementioned normal-mode-rejection means performs It is the normal-mode-rejection technique by the SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method. in that case the aforementioned sound model group A normal mode rejection is performed using the normal-mode-rejection technique by the SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method to each voice data superimposed on the noise from which the aforementioned kind differs. A feature vector is obtained from each voice data by which the normal mode rejection was carried out, and it is created using the feature vector. And a sound model group storage means to memorize the sound model group created by this, The 1st voice feature analysis means which performs the 1st voice feature analysis for obtaining the feature analytical data in a frequency domain to the voice data for recognition superimposed on the aforementioned noise, When the noise section or the voice section is judged and it judges with it being the noise section using the feature analysis result The noise section / a voice section judging means to save the feature data of the noise section for the feature data-storage means, and when it is judged with it being the voice section A noise kind judging means to judge the kind of noise on which it is superimposed with the feature data by which preservation was carried out [aforementioned], A sound model-selection means to choose a predetermined sound model based on the judgment result out of the aforementioned sound model group prepared for kind correspondence of the aforementioned noise, A normal-mode-rejection means to perform a normal mode rejection to the voice data for recognition superimposed on the aforementioned noise using the normal-mode-rejection technique by the describing [above] SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method, It is considering as composition with the 2nd voice feature analysis means which performs the 2nd voice feature analysis for obtaining the feature data required for speech recognition, and a speech recognition means to perform speech recognition to the feature analysis result using the sound model by which selection was carried out [aforementioned], to the voice data by which the normal mode rejection was carried out.

[0027] Moreover, the normal-mode-rejection technique which the aforementioned normal-mode-rejection means performs is the normal-mode-rejection technique by the cepstrum average normalizing method, and in that case, to each voice data superimposed on the noise from which the aforementioned kind differs, the aforementioned sound model performs a normal mode rejection using the normal-mode-rejection technique by the cepstrum average normalizing method, and is created using the feature vector of each voice data obtained by it. And a sound model group storage means to memorize the sound model group created by this, A feature analysis means to perform the feature analysis for asking for the feature vector showing a cepstrum coefficient from the voice data for recognition superimposed on the aforementioned noise, When the noise section or the voice section is judged and it judges with it being the noise section using the feature analysis result When it judges with the feature vector of the noise section being saved for the feature data-storage means, and it being the voice section By the feature vector of the noise section saved for the noise section / a voice section judging means to save the feature vector of the voice section for the feature analytical-data storage means, and this feature data-storage means A noise kind judging means to judge the kind of noise on which it is superimposed, and a sound

carried out [aforementioned] to the feature vector.
[0031] In such a voice recognition unit the sound model of noise correspondence of each above In addition to the kind of noise, it considers as the sound model also corresponding to the S/N ratio of two or more step story for every kind of each noise. the sound model in that case Each noise from which a S/N ratio differs for two or more kinds of every noises creates each voice data on which it was superimposed for every kind of noise. A normal mode rejection is performed to voice data using the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique, respectively. It is the sound model of S/N ratio correspondence of two or more step story per kind of each noise created to each of the voice data by which the normal mode rejection was carried out using the feature vector of each voice data obtained with the application of the cepstrum average normalizing method.
[0032] When a sound model is a sound model also corresponding to the S/N ratio of two or more step story per kind of each noise, and the aforementioned noise kind judging means It adds to the processing which judges the kind of noise to the voice data for recognition superimposed on the aforementioned noise. Processing which asks for a S/N ratio from the size of the noise of the noise section and the size of the voice of the voice section is performed, and the aforementioned sound model-selection section is made to choose a sound model based on the judged noise kind and the called-for S/N ratio.
[0033] Furthermore, the voice recognition unit of this invention creates each voice data superimposed on the noise from which a S/N ratio differs for every kind of noise. The sound model group corresponding to the S/N ratio which performed the normal mode rejection using the predetermined normal-mode-rejection technique, and was created by the feature vector of each of the voice data by which the normal mode rejection was carried out to each voice data of these, A sound model group storage means to memorize this sound model group, and a S/N ratio judging means to judge the S/N ratio of the noise on which it is superimposed to the voice data for recognition superimposed on noise, A sound model group-selection means to choose a predetermined sound model out of the sound model group corresponding to the aforementioned S/N ratio based on the judgment result, A normal-mode-rejection means to perform a normal mode rejection to the voice data for recognition superimposed on the aforementioned noise using the aforementioned predetermined normal-mode-rejection method, You may consider as composition with a speech recognition means to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of the voice data by which the normal mode rejection was carried out.
[0034] The normal-mode-rejection technique in that case may be the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique, and may be the normal-mode-rejection technique by the cepstrum average normalizing method.
[0035] Thus, this invention creates each voice data superimposed on the noise from which a kind differs, and creates the sound model corresponding to [perform / as opposed to / the voice data of ** / respectively] a normal mode rejection using the predetermined normal-mode-rejection technique, and / the kind of noise using each of the voice data by which the normal mode rejection was carried out superimposed on each noise of these. And at the time of actual recognition, the kind of noise on which it is superimposed is judged to the voice data for recognition superimposed on noise. While choosing a predetermined sound model out of the sound model corresponding to the kind of the aforementioned noise based on the judgment result To the voice data for recognition superimposed on the aforementioned noise, a normal mode rejection is performed using the aforementioned predetermined normal-mode-rejection method, and it is made to perform speech recognition to the voice data by which the normal mode rejection was carried out using the sound model by which selection was carried out [aforementioned].
[0036] By this, the recognition processing using the optimal sound model according to the kind of noise on which it is superimposed is attained, and even if it is under the environment where predetermined noise exists, a high recognition rate can be obtained.
[0037] When 2 or 3 kinds of noises exist in the operating environment of a device regularly

[0043] moreover It is also possible to create the sound model using both the SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method, and the cepstrum average normalizing method. In this case, when performing actual speech recognition, after performing the normal mode rejection using the SUPEKUTORARU subtraction method or the continuation SUPEKUTORARU subtraction method Since a feature vector is generated by the cepstrum average normalizing method and it is made to pass it to the speech recognition section as a feature vector for speech recognition to the voice data by which the normal mode rejection was carried out the additivity noise which could obtain the still higher recognition performance and was mentioned above in this case, and multiplication -- the correspondence to broad noises, such as sex noise, is attained

[0044] this invention can still also perform speech recognition which took two or more S/Ns into consideration about the noise it was decided that a certain specification would be. In this case, each voice data on which it was superimposed for every S/N ratio is created, and to each voice data of these, a certain specific kind from which a S/N ratio differs of noise performs a normal mode rejection using the predetermined normal-mode-rejection technique, and creates the sound model group corresponding to each S/N ratio by using the feature vector of each of the voice data by which the normal mode rejection was carried out. And at the time of actual recognition, the voice data for recognition superimposed on noise is received. Judge the S/N ratio on which it is superimposed and a predetermined sound model is chosen out of the sound model group corresponding to each S/N ratio based on the judgment result. To the voice data for recognition superimposed on the aforementioned noise, a normal mode rejection is performed using the aforementioned predetermined normal-mode-rejection method, and it is made to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the feature vector of the voice data by which the normal mode rejection was carried out.

[0045] Even if it can specify the kind of noise, when performing speech recognition under environment with changing [much] the size (S/N ratio), it will become convenient, and this can make the recognition rate under such environment high.

[0046]

[Embodiments of the Invention] Hereafter, the form of operation of this invention is explained. In addition, the contents explained with the form of this operation also include the concrete contents of processing of the speech recognition processing program in the record medium which recorded the speech recognition processing program of this invention while being the speech recognition method of this invention, and explanation about a voice recognition unit.

[0047] Although speech recognition is performed to the voice data from which this invention removed fundamentally the noise superimposed on the voice used as a processing object, and noise was removed The sound model used for the speech recognition assumes some kinds of noise (steady noise). The voice data which was made to superimpose each noise on the voice data (clean voice data of noise which is not) to a certain voice, and was superimposed on noise is generated. Processing which removes noise is performed from the voice data superimposed on the noise, and a sound model is created using the voice wave after the normal-mode-rejection processing (it differs from clean voice data without noise somewhat).

[0048] That is, the sound model from which the noise was removed for every kind of noise in the procedure mentioned above will be created for every kind of some noises prepared beforehand.

[0049] And in performing actual speech recognition, while judging the kind of noise on which the voice data for recognition is overlapped, processing which removes the noise is performed, a sound model is chosen according to the kind of noise, and speech recognition processing is performed using the selected sound model.

[0050] Furthermore, the sound model which set the S/N ratio which is a ratio of the size of noise and voice data as some with the kind of each noise of these is created. For example, although three kinds of sound models are created when only the kind of these noises is taken into consideration if three kinds, noise N1, noise N2, and noise N3, were chosen for the kind of noise If two steps of S/N ratios are taken into consideration about each noise, since processing which set up two kinds of sizes of noise and mentioned above will be performed about each

section for n frames and the 1st voice feature analysis processing which were memorized by the feature data-storage section 5 in Step s3. Each feature data of these shows audio power, while the kind and power of noise are known, since power etc. is obtained other than the feature of a frequency component.

[0060] For example, with the form of this 1st operation, steady noises, such as bustle in the run sound of an automobile, the operation sound of an air-conditioner, and a town, are assumed as noise. Here, three kinds shall be considered as such steady noise, and it shall be expressed with noise N1, noise N2, and noise N3. Therefore, it can judge whether it is close to which of noises N1, N2, and N3 by investigating the feature data for n frames of the noise section.

[0061] Moreover, if the power of noise and audio power are known, it can ask for a S/N ratio. In addition, since it is necessary to calculate a S/N ratio in the place where the power of the voice section had a certain amount of size in order to ask for a S/N ratio, a S/N ratio is calculated using the maximum and the average for several frames or all frames in the voice section.

[0062] Thus, if a S/N ratio is called for while the kind of noise is judged next, sound model-selection operation will be performed. With the form of this 1st operation, as for the sound model, the value of a S/N ratio has prepared the sound model below L1, and the sound model beyond L1 to these three kinds of noises N1, N2, and N3 supposing three kinds of steady noises N1, N2, and N3.

[0063] for example, with the form of this 1st operation When a S/N ratio is less than [L1], the kind of noise with noise N1 The sound model M1, When a S/N ratio is more than L1 with noise N1, a S/N ratio is less than [L1] with the sound model M2 and noise N2, a S/N ratio is more than L1 with the sound model M3 and noise N2 and a S/N ratio is less than [L1] with the sound model M4 and noise N3, the sound model M5, When a S/N ratio is more than L1 with noise N3, suppose that it is matched like the sound model M6. Therefore, in the sound model group storage section 7, these six kinds of sound models M1, M2, ..., M6 are saved. These sound models M1, M2, ..., M6 are created as follows.

[0064] That is, the noise of six patterns which have two steps of S/N ratios (is it less than [L1] or more than L1?) about noises N1, N2, and N3 and each noise is prepared, and the voice data of six patterns is created by making the noise of these 6 pattern superimpose on voice data without noise.

[0065] The voice data by which the voice data of these six patterns was superimposed in the S/N ratio on the noise N1 below L1, Voice data superimposed in the S/N ratio on the noise N1 beyond L1, voice data superimposed in the S/N ratio on the noise N2 below L1, Voice data superimposed in the S/N ratio on the noise N2 beyond L1, S/N ratios are the voice data superimposed on the noise N3 below L1, and the voice data of six patterns of the voice data superimposed in the S/N ratio on the noise N3 beyond L1.

[0066] To each voice data of these 6 pattern, a normal mode rejection is performed using the predetermined normal-mode-rejection technique, and six kinds of sound models M1, M2, ..., M6 are created by using the feature vector obtained by carrying out feature-analysis processing of the voice data of the six patterns by which the normal mode rejection was carried out.

[0067] Here, in the processing in Step s5, it judges that the kind of noise is close to noise N1. When the called-for S/N ratio is <(S/N) L1, i.e., less than [L1], the sound model M1 is chosen from the sound model group storage section 7.

[0068] Thus, selection of the sound model according to the kind and S/N ratio of a noise makes [next] the normal-mode-rejection processing by the normal-mode-rejection section 8 (Step s6). With the gestalt of this 1st operation, this normal-mode-rejection processing is normal-mode-rejection processing by the SS method or the CSS method, and performs SUPEKUTORARU subtraction using the feature data of the noise section for n newest frames memorized by the feature data-storage section 5 in Step s3 mentioned above, and the feature data of the voice section. The voice data from which noise was removed can be obtained by this. However, such after normal-mode-rejection processing was carried out, even if it was, it was left behind to voice data although the influence of noise was slight.

this SS method has an effect in normal mode rejections, such as bustle in the run sound of an automobile, the operation sound of an air-conditioner, and a town, a big effect is acquired by being applied to the device used in many cases under environment with much such noise.

[0079] [Gestalt of the 2nd operation] The cepstrum average normalizing method (the CMN method) is used for the gestalt of the 2nd operation as the normal-mode-rejection method. If drawing 3 is drawing showing the outline composition of the voice recognition unit of the gestalt of this 2nd operation and only components are enumerated A microphone 1, amplifier, and an A/D converter It has composition with the input speech processing section 2 which it has, the voice feature-analysis section 21, the noise section / voice section judging section 4, the feature data-storage section 5, the noise kind judging / sound model-selection section 6, the sound model group storage section 7, the normal-mode-rejection section 8, the speech recognition section 10, the language model storage section 11, etc. The explanation of operation which referred to the flow chart of drawing 4 explains the function of each [these] component serially.

[0080] in drawing 4 , the voice feature-analysis section 21 performs a voice feature analysis to the processing-object voice data after A/D conversion first for every (the time length of one frame is about about twenty msecs) frame (Step s21) This voice feature analysis presupposes that it is a feature analysis for asking for a cepstrum coefficient (for example, a mel cepstrum coefficient and an LPC cepstrum coefficient) with the gestalt of this 2nd operation.

[0081] And based on the voice feature-analysis result, it judges whether it is the noise section or it is the voice section by the noise section / voice section judging section 4 (Step s22). When judged with it being the noise section, this noise section / voice section judging section 4 judge further whether it is the noise section where the noise section exists ahead [of the voice section / direction of time-axis], or it is the noise section which exists behind [direction of time-axis] the voice section (Step s23).

[0082] When it is the noise section which exists ahead [of the voice section / direction of time-axis] as a result of this judgment, the feature data-storage section 5 is made to memorize the feature data for the n1 newest frame which the feature analysis was carried out and was obtained (feature vector of a cepstrum coefficient) (Step s24).

[0083] moreover, n which constitutes the voice section when it judges that the result which judged whether it is the noise section or it was the voice section is the voice section (from the start of the voice section to an end) -- the feature data for two frames (feature vector of a cepstrum coefficient) are memorized in the feature data-storage section 5 (Step s25)

[0084] Furthermore repeat a voice feature analysis and the judged result whether it is the noise section or it is the voice section As what (Steps s21, s22, and s23) and the voice section ended when judged with it being the noise section where it is judged with it being the noise section, and the noise section exists behind [direction of time-axis] the voice section n after a voice section end -- the feature data for three frames (feature vector of a cepstrum coefficient) are memorized in the feature data-storage section 5 (Step s26)

[0085] and this n -- if it judged whether the storage processing for three frames was completed (Step s27) and processing is completed, it will go into noise kind judging operation and sound model-selection operation by a noise kind judging / sound model-selection section 6 (Step s28) This noise kind judging operation and sound model-selection operation are explained below.

[0086] n1 and n which boil the kind of this noise, the judgment of a size (S/N ratio), and sound model-selection operation till then, and are memorized by the feature data-storage section 5 -- it carries out using each feature data for two frames

[0087] That is, the kind of noise can judge whether noise is close to which noise using the feature data (for example, n the feature data for one frame) of the noise section, and it can ask for a S/N ratio with the size of power and the size of the power of the voice section which are obtained by carrying out the feature analysis of the noise section.

[0088] In addition, also in the gestalt of this 2nd operation, processing corresponding to three kinds of noises N1, N2, and N3 shall be performed.

[0089] And based on the judgment result of the kind of these noises, and the size of the called-for S/N ratio, sound model-selection operation of which sound model to use is performed. it judges with this sound model-selection operation having the kind of noise close

every frame for 20 frames which constitute the voice section, this re-calculation subtracts the average feature vector Cm, and it performs C1'=C1-Cm, C2'=C2-Cm, ..., C20'=C20-Cm in this example. and -- asking -- having had -- C -- one -- ' -- C -- one -- ' ... C -- 20 -- ' -- the feature vector for 20 frames after normal-mode-rejection processing -- becoming .

[0102] this -- a feature vector -- C -- one -- ' -- C -- one -- ' ... C -- 20 -- ' -- speech recognition -- the section -- ten -- giving -- having -- speech recognition -- the section -- ten -- **** -- choosing -- having had -- sound -- a model -- beforehand -- preparing -- having -- **** -- language -- a model -- 11 -- having used -- speech recognition -- processing -- carrying out .

[0103] Thus, it is made to perform speech recognition using the sound model which the sound model mentioned above also in the form of the 2nd operation corresponding to the kind of noise and the size of a S/N ratio like the form of the 1st operation was chosen, and was chosen, and the language model saved in the language model storage section 11.

[0104] In addition, six kinds of sound models used with the form of this 2nd operation Three kinds of noises N1, N2, and N3 which have two steps of S/N ratios are made to superimpose on voice data (clean voice data of noise which is not) like the form of the 1st operation. Generate the voice data of six patterns with which it was superimposed on noise, and processing (normal-mode-rejection processing by the CMN method) which removes noise to the voice data of the six patterns, respectively is performed. It is created using the voice data (voice data to which the influence of noise was left behind somewhat unlike clean voice data without noise) of six patterns after the normal-mode-rejection processing. That is, it can be said that it is the sound model created by the voice data near the voice data used as an actual speech recognition processing object.

[0105] Therefore, a still higher recognition performance can be obtained by choosing the optimal sound model and performing speech recognition to the voice data used as an actual speech recognition processing object, using the selected sound model based on the kind of noise and the size of a S/N ratio on which the voice data is overlapped.

[0106] Moreover, the CMN method as a normal-mode-rejection method of the form of this 2nd operation can perform a normal mode rejection in the few amount of operations, also by low CPU of arithmetic proficiency, can respond enough and becomes realizable [on small-scale and cheap hardware]. moreover, since this CMN method has an effect in removal of the noise (multiplication sex noise) originating in space transfer characteristics, such as a property, an echo, etc. of a microphone, a big effect is acquired by being applied to the device used in many cases under the environment which such noise tends to generate

[0107] [Form of the 3rd operation] The form of this 3rd operation combines the form of the 1st operation, and the form of the 2nd operation. Also in the form of this 3rd operation, like the form of the 1st and the 2nd operation, although six sound models M1, M2, ..., M6 shall be prepared according to the kind of noise, and the size of a S/N ratio, the sound model used in the form of this 3rd operation is created as follows.

[0108] Three kinds of noises N1, N2, and N3 which have two steps of S/N ratios are made to superimpose on voice data (clean voice data of noise which is not), as mentioned above. Generate the voice data of six patterns with which it was superimposed on noise, and processing (normal-mode-rejection processing by the SS method or the CSS method) which removes noise to the voice data of the six patterns, respectively is performed. The voice data (voice data to which the influence of noise was left behind somewhat unlike clean voice data without noise) of six patterns after the normal-mode-rejection processing is generated.

[0109] And the CMN method is applied to each voice section of the voice data of six patterns by which the normal mode rejection was carried out by this SS method or the CSS method. namely, the feature vector (n feature vector for two frames) obtained by carrying out the feature analysis of the voice section in each voice data as mentioned above -- using -- the n -- it asks for the feature vector of the average for two frames For example, it considers as n2=20, then the average Cm of the feature vector for 20 frames (expressing this with C1, C2, ..., C20, each [these] feature vectors C1, C2, ..., C20 have it it, for example, a 10-dimensional component).

[0110] Next, the feature vector of the voice section (here 20 frames) is re-calculated using the feature vector of the called-for average. That is, C1'=C1-Cm, C2'=C2-Cm, ..., C20'=C20-Cm

[0119] After this sound model-selection processing is completed next, the feature data generation processing for obtaining the voice feature data required performing speech recognition is performed by the CMN operation part 31 (Steps s49 and s50). This feature data generation processing is performed using the CMN method as a normal-mode-rejection method mentioned above.
[0120] the feature vector (n feature vector for two frames) according to the feature analysis result of the voice section as the form of the 2nd operation explained this CMN method -- using -- the n -- it asks for the feature vector of the average for two frames in the same procedure as the above-mentioned (the feature vector of the called-for average is set to Cm) The feature vector of the voice section (here 20 frames) is re-calculated using the feature vector Cm of this average. That is, C1'=C1-Cm, C2'=C2-Cm, ..., C20'=C20-Cm are performed.
[0121] and -- asking -- having had -- C -- one -- ' -- C -- one -- ' ... C -- 20 -- ' -- obtaining -- having had -- 20 -- a frame -- a part -- each -- a frame -- every -- a feature vector -- becoming . and -- this -- each -- a frame -- every -- a feature vector -- C -- one -- ' -- C -- one -- ' ... C -- 20 -- ' -- speech recognition -- the section -- ten -- giving -- having -- speech recognition -- the section -- ten -- **** -- choosing -- having had -- sound -- a model -- language -- a model -- storage -- the section -- 11 -- saving -- having -- **** -- language -- a model -- having used -- speech recognition -- processing -- carrying out .
[0122] Thus, like the form of the 1st mentioned above also in the form of the 3rd operation, and the 2nd operation, the sound model according to the kind of noise and the size of a S/N ratio is chosen, and it is made to perform speech recognition using the selected sound model and the language model currently prepared beforehand.
[0123] With the form of this 3rd operation, the sound model which used both the SS method (or the CSS method) and the CMN method is created. When performing actual speech recognition, after performing the normal mode rejection using the SS method (or the CSS method) Since a feature vector is generated by the CMN method and it is made to pass it to the speech recognition section 10 as a feature vector for speech recognition to the voice data by which the normal mode rejection was carried out a still higher recognition performance -- it can obtain -- moreover -- the form of this 3rd operation -- additivity noise and multiplication -- the correspondence to broad noises, such as sex noise, is attained
[0124] In addition, this invention is not limited to the form of the operation explained above, and the deformation implementation of it is variously attained in the range which does not deviate from the summary of this invention. For example, although the form of each above-mentioned operation showed the example which made the kind of noise three kinds, noise N1, noise N2, and noise N3, and the S/N ratio made two steps of sizes about each [these] noise, it is not restricted to this.
[0125] Especially the kind of noise does not consider each as independent noise, but you may make it consider as one noise what combined some noises like the bustle in the run sound of an automobile, the operation sound of an air-conditioner, and a town.
[0126] After generating the voice data which made the run sound of an automobile, and the operation sound of an air-conditioner superimpose on the voice data recorded under environment without noise as an example simultaneously and removing noise from this generated voice data using the predetermined normal-mode-rejection method, the sound model for speech recognition learned using the voice data from which the noise was removed can also be created.
[0127] Thus, since it is possible to create arbitrarily the sound model created combining the stationary noise which is easy to exist under the environment where a device is used two or more kinds, a still higher recognition rate can be obtained by preparing some optimal sound models for each device correspondence. Furthermore, if that from which a S/N ratio differs is created about each noise of these, a good result will be obtained more.
[0128] Moreover, it is not necessary to constitute the composition of the voice recognition unit shown by drawing 1 , drawing 3 , and drawing 5 as the example of the form of operation is shown, respectively and these drawings showed. For example, although a means to judge a noise kind, and a means to choose a sound model should be summarized to one as a noise kind judging means / a sound model-selection means 6, of course, you may make it prepare each as

[0137] Moreover, this invention can create the processing program with which the procedure for realizing this invention explained above was described, the processing program can be made to record on record media, such as a floppy disk, an optical disk, and a hard disk, and this invention also contains the record medium with which the processing program was recorded. Moreover, you may make it obtain the processing program concerned from a network.

[0138]

[Effect of the Invention] As explained above, this invention creates each voice data superimposed on the noise from which a kind differs, and creates the sound model corresponding to [perform / as opposed to / the voice data of ** / respectively] a normal mode rejection using the predetermined normal-mode-rejection technique, and / the kind of noise using each of the voice data by which the normal mode rejection was carried out superimposed on each noise of these. And at the time of actual recognition, the kind of noise on which it is superimposed is judged to the voice data for recognition superimposed on noise. While choosing a predetermined sound model out of the sound model corresponding to the kind of the aforementioned noise based on the judgment result To the voice data for recognition superimposed on the aforementioned noise, a normal mode rejection is performed using the aforementioned predetermined normal-mode-rejection method, and it is made to perform speech recognition to the voice data by which the normal mode rejection was carried out using the sound model by which selection was carried out [aforementioned].

[0139] By this, the recognition processing using the optimal sound model according to the kind of noise on which it is superimposed is attained, and even if it is under the environment where predetermined noise exists, a high recognition rate can be obtained.

[0140] When 2 or 3 kinds of noises exist in the operating environment of a device regularly especially, the sound model for every noises of those is created, and a high recognition rate can be realized by performing speech recognition processing which was mentioned above using the sound model.

[0141] And as one of the normal-mode-rejection technique used by this invention, it is the normal-mode-rejection technique by SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique, and the normal mode rejection at the time of the aforementioned sound model creation is performed in that case using the SUPEKUTORARU substraction method or continuation SUPEKUTORARU subtraction technique. Moreover, after judging the kind of noise on which it is superimposed, while choosing the optimal sound model with the feature-analysis data of the noise section based on the judgment result at the time of actual recognition A normal mode rejection is performed to the voice data for recognition superimposed on the aforementioned noise using the normal-mode-rejection technique by the spectrum substraction method. It is made to perform speech recognition using the sound model by which selection was carried out [aforementioned] to the result obtained by carrying out the feature analysis of the voice data by which the normal mode rejection was carried out.

[0142] Thus, by using SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique as the normal-mode-rejection method, normal-mode-rejection processing can be performed in the few amount of operations, and it can respond enough also by low CPU of arithmetic proficiency. It becomes realizable [on small-scale and cheap hardware] by this. moreover, since this SUPEKUTORARU subtraction technique or continuation SUPEKUTORARU subtraction technique has an effect in removal of noises (generally called additivity noise), such as bustle in the run sound of an automobile, the operation sound of an air-conditioner, and a town, a big effect is acquired by being applied to the device used in many cases under environment with much such noise

[0143] Moreover, also using the normal-mode-rejection technique by the cepstrum average normalizing method as other examples of the normal-mode-rejection technique and a thing are made. In this case, the normal mode rejection at the time of the aforementioned sound model creation is performed using the cepstrum average normalizing method. Moreover, after judging the kind of noise on which it is superimposed, while choosing the optimal sound model with the feature-analysis data of the noise section based on the judgment result at the time of

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is a block diagram for explaining the gestalt of operation of the 1st of the voice recognition unit of this invention.
[Drawing 2] It is a flow chart for explaining the procedure of the gestalt of the 1st operation.
[Drawing 3] It is a block diagram for explaining the gestalt of operation of the 2nd of the voice recognition unit of this invention.
[Drawing 4] It is a flow chart for explaining the procedure of the gestalt of the 2nd operation.
[Drawing 5] It is a block diagram for explaining the gestalt of operation of the 3rd of the voice recognition unit of this invention.
[Drawing 6] It is a flow chart for explaining the procedure of the gestalt of the 3rd operation.
[Description of Notations]
1 Voice Input Section
2 Input Speech Processing Section
3 1st Voice Feature-Analysis Section
4 Noise Section / Voice Section Judging Section
5 The Feature Data-Storage Section
6 Noise Kind Judging / Sound Model-Selection Section
7 Sound Model Group Storage Section
8 Normal-Mode-Rejection Section
9 2nd Voice Feature-Analysis Section
10 Speech Recognition Section
11 Language Model Storage Section
21 Feature-Analysis Section
31 CNN Operation Part (Normal-Mode-Rejection Section by the CNN Method)

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－132289
（P2002－132289A）
(43)公開日　平成14年５月９日(2002.5.9)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ１０Ｌ　15/20		Ｇ１０Ｌ　3/00	５３１Ｑ	５Ｄ０１５
15/06			５２１Ｔ	
21/02		3/02	３０１Ｄ	

審査請求　未請求　請求項の数33　ＯＬ　（全 25 頁）

(21)出願番号　特願2000－322914(P2000－322914)

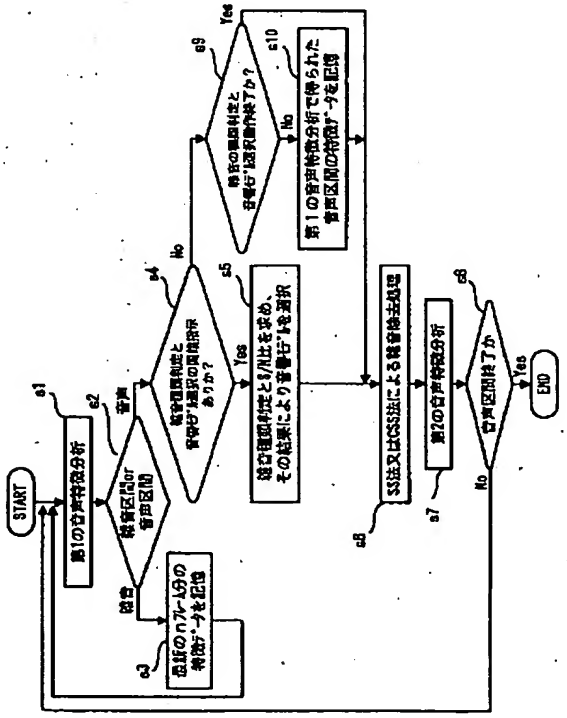
(22)出願日　平成12年10月23日(2000.10.23)

(71)出願人　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号
(72)発明者　宮沢　康永
長野県諏訪市大和３丁目３番５号　セイコーエプソン株式会社内
(74)代理人　100095728
弁理士　上柳　雅誉　（外１名）

Ｆターム(参考)　5D015 AA05 CC11 EE05

(54)【発明の名称】　音声認識方法および音声認識処理プログラムを記録した記録媒体ならびに音声認識装置

(57)【要約】
【課題】複数種の雑音の存在する環境下であっても高い認識率を得るようにする。
【解決手段】種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクションの雑音除去手法（ＳＳ法という）を用いて雑音除去を行い、雑音除去後のそれぞれの音声データを特徴分析して得られた特徴ベクトルにより作成された雑音の種類対応の音響モデルを用意する。認識時には、認識対象音声データに対し、第１の音声特徴分析を行い、雑音区間／音声区間を判定し、雑音区間の場合にはその特徴データを保存し（ステップｓ１～ｓ３）、音声区間の場合には保存された特徴データによって雑音種類を判定し、その結果によって、対応する音響モデルを選択する（ステップｓ５）。そして、認識対象音声データに対し、ＳＳ法による雑音除去を行い、その雑音除去後の音声データに対し、音声認識用の特徴ベクトルを得るための第２の音声特徴分析を行う（ステップｓ６，ｓ７）。



1

【特許請求の範囲】

【請求項１】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、雑音の種類に対応する音響モデル群を作成してそれを保持しておき、

認識時には、

雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定し、その判定結果に基づいて、前記雑音の種類に対応した音響モデル群の中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うことを特徴とする音声認識方法。

【請求項２】　前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル群は、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、雑音除去されたそれぞれの音声データから特徴ベクトルを得て、その特徴ベクトルを用いて作成され、

認識時には、

前記雑音が重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行い、

その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存し、音声区間であると判定された場合には、前記保存された特徴データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択し、

前記雑音が重畳された認識対象音声データに対し、前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、

その雑音除去された音声データに対し、音声認識に必要な特徴データを得るための第２の音声特徴分析を行い、その特徴分析結果に対し、前記選択された音響モデルを用いて音声認識を行う、

ことを特徴とする請求項１記載の音声認識方法。

【請求項３】　前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であって、その場合、前記音響モデルは、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去を行い、それによって得られたそれぞれの音声データの特徴ベクトルを用いて作成され、

2

認識時には、

前記雑音が重畳された認識対象音声データに対し、ケプストラム係数を表す特徴ベクトルを求めるための第１の音声特徴分析を行い、

その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その特徴ベクトルを保存し、音声区間であると判定された場合には、その音声区間の開始から終了までの音声区間に対応する特徴ベクトルを保存し、前記保存された雑音区間の特徴ベクトルによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択し、

前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去処理を行い、

その雑音除去処理後の特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う、

ことを特徴とする請求項１記載の音声認識方法。

【請求項４】　前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、このそれぞれ音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比に対応した音響モデルであることを特徴とする請求項１から３のいずれかに記載の音声認識方法。

【請求項５】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項４記載の音声認識方法。

【請求項６】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトル

3

によって、それぞれの雑音対応の音響モデル群を作成してそれを保存しておき、
認識時には、
雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行い、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存し、
音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、
その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の音声特徴分析処理を行い、その音声区間の特徴ベクトルを保存し、
音声区間終了後に、前記保存された雑音区間の特徴データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音対応に用意された音響モデル群の中から所定の音響モデルを選択し、
前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルに対し、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、
そのケプストラム平均正規化法を適用して得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う、
ことを特徴とする音声認識方法。

【請求項７】　前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音を雑音の種類ごとに重畳させたそれぞれの音声データを作成し、この音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対し、ケプストラム平均正規化法を適用して得られたそれぞれの音声データの特徴ベクトルを用いて作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比対応の音響モデルであることを特徴とする請求項６記載の音声認識方法。

【請求項８】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項７記載の音声認識方法。

4

【請求項９】　Ｓ／Ｎ比の異なるある特定の種類の雑音がそれぞれのＳ／Ｎ比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのＳ／Ｎ比に対応する音響モデル群を作成してそれを保持しておき、
認識時には、
雑音が重畳された認識対象音声データに対し、重畳されているＳ／Ｎ比を判定し、その判定結果に基づいて、前記Ｓ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うことを特徴とする音声認識方法。

【請求項１０】　前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であることを特徴とする請求項９記載の音声認識方法。

【請求項１１】　前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であることを特徴とする請求項９記載の音声認識方法。

【請求項１２】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データを特徴分析処理して得られた特徴ベクトルによって、雑音の種類に対応する音響モデル群を作成して、それを音響モデル群記憶手段に記憶させる手順と、
雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定し、その判定結果に基づいて、前記音響モデル群記憶手段に記憶された音響モデル群の中から所定の音響モデルを選択する手順と、
前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う手順と、
その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う手順と、
をその処理プログラムに含むことを特徴とする音声認識処理プログラムを記録した記録媒体。

【請求項１３】　前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル群は、種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、雑音除去されたそれぞれの音声データから特徴ベクトルを得て、そ

5

の特徴ベクトルを用いて作成され、
認識時の処理手順は、
前記雑音が重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行う手順と、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存し、音声区間であると判定された場合には、前記保存された特徴データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択する手順と、
前記雑音が重畳された認識対象音声データに対し、前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行う手順と、
その雑音除去された音声データに対し、音声認識に必要な特徴データを得るための第２の音声特徴分析を行い、その特徴分析結果に対し、前記選択された音響モデルを用いて音声認識を行う手順と、
を含むことを特徴とする請求項１２記載の音声認識処理プログラムを記録した記録媒体。

【請求項１４】　前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であって、その場合、前記前記音響モデルは、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去を行い、それによって得られたそれぞれの音声データの特徴ベクトルを用いて作成され、
認識時の処理手順は、
前記雑音が重畳された認識対象音声データに対し、ケプストラム係数を表す特徴ベクトルを求めるための第１の音声特徴分析を行う手順と、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その特徴ベクトルを保存し、音声区間であると判定された場合には、その音声区間の開始から終了までの音声区間に対応する特徴ベクトルを保存し、前記保存された雑音区間の特徴ベクトルによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択する手順と、
前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去処理を行う手順と、
その雑音除去処理後の特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う手順と、
を含むことを特徴とする請求項１２記載の音声認識処理プログラムを記録した記録媒体。

6

【請求項１５】　前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、このそれぞれ音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比に対応した音響モデルであることを特徴とする請求項１２から１４のいずれかに記載の音声認識処理プログラムを記録した記録媒体。

【請求項１６】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項１５記載の音声認識処理プログラムを記録した記録媒体。

【請求項１７】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトルによって、それぞれの雑音対応の音響モデル群を作成し、それを音響モデル群記憶手段に記憶させる手順と、
雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行う手順と、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存する手順と、
音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行う手順と、
その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の音声特徴分析処理を行い、その音声区間の特徴ベクトルを保存する手順と、
音声区間終了後に、前記保存された雑音区間の特徴分析データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音対応に用意された音響モデル群の中から所定の音響モデルを選択する手順と、

前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルに対し、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得る手順と、

そのケプストラム平均正規化法を適用して得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う手順と、

をその処理プログラムに含むことを特徴とする音声認識処理プログラムを記録した記録媒体。

【請求項１８】　前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音を雑音の種類ごとに重畳させたそれぞれの音声データを作成し、この音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対し、ケプストラム平均正規化法を適用して得られたそれぞれの音声データの特徴ベクトルを用いて作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比対応の音響モデルであることを特徴とする請求項１７記載の音声認識処理プログラムを記録した記録媒体。

【請求項１９】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項１８記載の音声認識処理プログラムを記録した記録媒体。

【請求項２０】　Ｓ／Ｎ比の異なるある特定の種類の雑音がそれぞれのＳ／Ｎ比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのＳ／Ｎ比に対応する音響モデル群を作成して、それを音響モデル群記憶手段に記憶させる手順と、

雑音が重畳された認識対象音声データに対し、重畳されているＳ／Ｎ比を判定し、その判定結果に基づいて、前記Ｓ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択する手順と、

前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う手順と、

その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う手順と、

を処理プログラムに含むことを特徴とする音声認識処理プログラムを記録した記録媒体。

【請求項２１】　前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であることを特徴とする請求項２０記載の音声認識処理プログラムを記録した記録媒体。

【請求項２２】　前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であることを特徴とする請求項２０記載の音声認識処理プログラムを記録した記録媒体。

【請求項２３】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成された雑音の種類に対応する音響モデル群と、

この音響モデル群を記憶する音響モデル群記憶手段と、

雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定する雑音判定手段と、

その判定結果に基づいて、前記雑音の種類に対応した音響モデル群の中から所定の音響モデルを選択する音響モデル群選択手段と、

前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う雑音除去手段と、

その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う音声認識手段と、

を有したことを特徴とする音声認識装置。

【請求項２４】　前記雑音除去手段が行う雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル群は、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、雑音除去されたそれぞれの音声データから特徴ベクトルを得て、その特徴ベクトルを用いて作成され、

これによって作成された音響モデル群を記憶する音響モデル群記憶手段と、

前記雑音が重畳された認識対象音声データに対し、周波数領域での特徴分析データを得るための第１の音声特徴分析を行う第１の音声特徴分析手段と、

その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定した場合には、その雑音区間の特徴データを特徴データ記憶手段に保存する雑音区間／音声区間判定手段と、

音声区間であると判定された場合には、前記保存された

9

特徴データによって、重畳されている雑音の種類を判定する雑音種類判定手段と、
その判定結果に基づいて、前記雑音の種類対応に用意された前記音響モデル群の中から所定の音響モデルを選択する音響モデル選択手段と、
前記雑音が重畳された認識対象音声データに対し、前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声データに対し、音声認識に必要な特徴データを得るための第２の音声特徴分析を行う第２の音声特徴分析手段と、
その特徴分析結果に対し、前記選択された音響モデルを用いて音声認識を行う音声認識手段と、
を有したことを特徴とする請求項２３記載の音声認識装置。

【請求項２５】　前記雑音除去手段が行う雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であって、その場合、前記音響モデルは、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去を行い、それによって得られたそれぞれの音声データの特徴ベクトルを用いて作成され、
これによって作成された音響モデル群を記憶する音響モデル群記憶手段と、
前記雑音が重畳された認識対象音声データに対し、ケプストラム係数を表す特徴ベクトルを求めるための特徴分析を行う特徴分析手段と、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定した場合には、その雑音区間の特徴ベクトルを特徴データ記憶手段に保存し、音声区間であると判定した場合には、その音声区間の特徴ベクトルを特徴分析データ記憶手段に保存する雑音区間／音声区間判定手段と、
この特徴データ記憶手段に保存された雑音区間の特徴ベクトルによって、重畳されている雑音の種類を判定する雑音種類判定手段と、
その判定結果に基づいて、前記雑音の種類対応に用意された前記音響モデル群の中から所定の音響モデルを選択する音響モデル選択手段と、
前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去処理を行う雑音除去手段と、
その雑音除去処理によって得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う音声認識手段と、
有することを特徴とする請求項２３記載の音声認識装置。

【請求項２６】　前記それぞれの雑音対応の音響モデル

10

は、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、このそれぞれ音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比に対応した音響モデルであることを特徴とする請求項２３から２５のいずれかに記載の音声認識装置。

【請求項２７】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音種類判定手段は、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、前記音響モデル選択部は、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項２６記載の音声認識装置。

【請求項２８】　種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトルによって作成されたそれぞれの雑音対応の音響モデル群と、
この音響モデル群を記憶する音響モデル群記憶手段と、
雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行う第１の音声特徴分析手段と、
その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その特徴データを特徴データ記憶手段に保存する雑音区間／音声区間判定手段と、
音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法によるノイズ除去手法を用いて雑音除去を行う雑音除去手段と、
その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の特徴分析処理を行いその音声区間の特徴ベクトルを特徴データ記憶手段に保存する第２の音声特徴分析部と、
音声区間終了後に、前記保存された雑音区間の特徴データによって、重畳されている雑音の種類を判定する雑音種類判定手段と、
その判定結果に基づいて、前記雑音対応に用意された音

11

響モデルの中から所定の音響モデルを選択する音響モデル選択手段と、
前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを出力するケプストラム平均正規化演算部と、
その特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う音声認識部と、
を有することを特徴とする音声認識装置。

【請求項２９】　前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、それぞれ音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対し、ケプストラム平均正規化法を適用して得られたそれぞれの音声データの特徴ベクトルを用いて作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比対応の音響モデルであることを特徴とする請求項２８記載の音声認識装置。

【請求項３０】　前記音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音種類判定手段は、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、前記音響モデル選択部は、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うことを特徴とする請求項２９記載の音声認識装置。

【請求項３１】　Ｓ／Ｎ比の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたＳ／Ｎ比に対応する音響モデル群と、
この音響モデル群を記憶する音響モデル群記憶手段と、
雑音が重畳された認識対象音声データに対し、重畳されている雑音のＳ／Ｎ比を判定するＳ／Ｎ比判定手段と、
その判定結果に基づいて、前記Ｓ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択する音響モデル群選択手段と、
前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う雑音除去手段と、
その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う音声認識手段と、

12

を有したことを特徴とする音声認識装置。

【請求項３２】　前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であることを特徴とする請求項３１記載の音声認識装置。

【請求項３３】　前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であることを特徴とする請求項３１記載の音声認識装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は種々の背景雑音が存在する環境下においても高い認識性能を可能とする音声認識方法および音声認識処理プログラムを記録した記録媒体ならびに音声認識装置に関する。

【０００２】

【従来の技術】近年、音声認識機能を搭載した機器が広く用いられるようになってきている。このような機器の使用環境は様々であり、雑音の多い環境下で使用せざるを得ない場合も多い。

【０００３】このような場合、当然のことながら雑音に対する対策を講じる必要が出てくる。雑音の一例としては、たとえば、自動車の走行音、エアーコンディショナ（エアコンという）の運転音などの定常的な雑音が身近なものとして考えられるが、これらの定常的な雑音の存在する環境下での音声認識を行う方法として、従来、以下に示すような音声認識方法が用いられている。

【０００４】その第１の例として、雑音のない環境下で収録した音声データに上述したような定常的な雑音から得られた雑音データを重畳させた音声データを生成し、この生成された音声データを用いて学習された音声認識用の音響モデルを作成し、その音響モデルを用いて音声認識を行う方法がある。

【０００５】また、第２の例としては、スペクトラル・サブトラクション（Spectral Subtraction）などの雑音除去方法を用いて音声認識を行う方法もある。この音声認識方法は、入力音声データから雑音成分を除去して、雑音の除去された音声データに対して音声認識を行うが、その場合でも、上述同様、雑音のない環境下で収録した音声データに定常的な雑音から得られた雑音データを重畳させた音声データを生成し、この生成された音声データからスペクトラル・サブトラクション法による雑音除去方法を用いて雑音を除去したのちに、その雑音の除去された音声データを用いて学習した音声認識用の音響モデルを作成しておき、その音響モデルを用いて音声認識を行うことがなされている。

【０００６】

【発明が解決しようとする課題】上述したような音声認識方法を採用することによって、何の対策も講じない場合に比べ、雑音環境下における認識性能の向上はある程度は可能となると考えられるが、まだまだ問題点も多

い。

【０００７】すなわち、定常的な雑音は、上述したような自動車の走行音、エアコンの運転音などの他にも、雑踏による雑音など様々な種類があり、それぞれが異なった性質を持っている。

【０００８】上述した従来の２つの例で述べた音響モデルは、音響モデルを学習する際、ある特定の雑音のみを用いて学習している場合が多い。たとえば、自動車の走行音を雑音データとして用い、その雑音データを音声データに重畳させた音声データを生成し、この生成された音声データからスペクトラル・サブトラクションによる雑音除去方法を用いて雑音を除去したのちに、その雑音の除去された音声データを用いて学習した音声認識用の音響モデルを作成する。

【０００９】このように、ある特定の雑音について作成された音響モデルを用いて音声認識を行えば、そのような雑音が存在する環境下では比較的好結果が得られるが、状況によっては、それ以外の雑音が存在することも十分あり得る。その場合の認識率は当然のことながら低下する。

【００１０】また、雑音の種類以外にも、実際に認識すべき音声信号と雑音信号の比であるS/N比の大きさによっても認識性能は異なってくる。

【００１１】そこで本発明は、雑音の種類やＳ／Ｎ比の大きさに対応した高い認識性能を得ることができ、しかも、演算能力の低いＣＰＵを用いた安価なハードウエアで実現可能とすることを目的としている。

【００１２】

【課題を解決するための手段】上述した目的を達成するために本発明の音声認識方法は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、雑音の種類に対応する音響モデル群を作成してそれを保持しておく。そして、認識時には、雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定し、その判定結果に基づいて、前記雑音の種類に対応した音響モデル群の中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うようにしている。

【００１３】また、本発明の音声認識処理プログラムを記録した記録媒体は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データを特徴分析処理して得られた特徴ベクトルによって、雑音の種類に対応する音響モデル群を作成して、それを音響モデル群記憶手段に記憶させる手順と、雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定し、その判定結果に基づいて、前記音響モデル群記憶手段に記憶された音響モデル群の中から所定の音響モデルを選択する手順と、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う手順と、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う手順とをその処理プログラムに含むものである。

【００１４】これら各発明において、前記雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル群は、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、雑音除去されたそれぞれの音声データから特徴ベクトルを得て、その特徴ベクトルを用いて作成されている。そして、認識時には、前記雑音が重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行い、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存し、音声区間であると判定された場合には、前記保存された特徴データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択し、前記雑音が重畳された認識対象音声データに対し、前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去された音声データに対し、音声認識に必要な特徴データを得るための第２の音声特徴分析を行い、その特徴分析結果に対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００１５】また、前記雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であってもよく、その場合、前記音響モデルは、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去を行い、それによって得られたそれぞれの音声データの特徴ベクトルを用いて作成されている。そして、認識時には、前記雑音が重畳された認識対象音声データに対し、ケプストラム係数を表す特徴ベクトルを求めるための第１の音声特徴分析を行い、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その特徴ベクトルを保存し、音声区間であると判定された場合には、その音声区間の開始から終了ま

15

での音声区間に対応する特徴ベクトルを保存し、前記保存された雑音区間の特徴ベクトルによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類対応に用意された音響モデル群の中から所定の音響モデルを選択し、　前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去処理を行い、その雑音除去処理後の特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００１６】さらに、前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとすることも可能で、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、このそれぞれ音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比に対応した音響モデルとしている。

【００１７】そして、音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うようにしている。

【００１８】また、本発明のもう一つの音声認識方法は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトルによって、それぞれの雑音対応の音響モデル群を作成してそれを保存しておく。そして、認識時には、雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行い、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存し、音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の音声特徴分析処理を行い、その音声区間

16

の特徴ベクトルを保存し、音声区間終了後に、前記保存された雑音区間の特徴データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音対応に用意された音響モデル群の中から所定の音響モデルを選択し、前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルに対し、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、そのケプストラム平均正規化法を適用して得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００１９】また、もう一つの音声認識処理プログラムを記録した記録媒体は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトルによって、それぞれの雑音対応の音響モデル群を作成し、それを音響モデル群記憶手段に記憶させる手順と、　雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行う手順と、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その分析された特徴データを保存する手順と、音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行う手順と、その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の音声特徴分析処理を行い、その音声区間の特徴ベクトルを保存する手順と、音声区間終了後に、前記保存された雑音区間の特徴分析データによって、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音対応に用意された音響モデル群の中から所定の音響モデルを選択する手順と、前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルに対し、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得る手順と、そのケプストラム平均正規化法を適用して得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う手順とをその処理プログラムに含むものである。

【００２０】これら音声認識方法および音声認識処理プログラムを記録した記録媒体において、前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑

音ごとにS／N比の異なるそれぞれの雑音を雑音の種類ごとに重畳させたそれぞれの音声データを作成し、この音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対し、ケプストラム平均正規化法を適用して得られたそれぞれの音声データの特徴ベクトルを用いて作成されたそれぞれの雑音の種類ごとに複数段階のS／N比対応の音響モデルとしている。

【００２１】そして、音響モデルがそれぞれの雑音の種類ごとに複数段階のS／N比にも対応した音響モデルである場合、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからS／N比を求める処理を行い、判定された雑音種類と求められたS／N比に基づいて音響モデルの選択を行うようにする。

【００２２】さらに、本発明の音声認識方法は、S／N比の異なるある特定の種類の雑音がそれぞれのS／N比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのS／N比に対応する音響モデル群を作成してそれを保持しておき、認識時には、　雑音が重畳された認識対象音声データに対し、重畳されているS／N比を判定し、その判定結果に基づいて、前記S／N比に対応した音響モデル群の中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うようにしたものでもよい。

【００２３】さらに、本発明の音声認識処理プログラムを記録した記録媒体は、　S／N比の異なるある特定の種類の雑音がそれぞれのS／N比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのS／N比に対応する音響モデル群を作成して、それを音響モデル群記憶手段に記憶させる手順と、雑音が重畳された認識対象音声データに対し、重畳されているS／N比を判定し、その判定結果に基づいて、前記S／N比に対応した音響モデル群の中から所定の音響モデルを選択する手順と、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う手順と、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う手順とを含んだ処理プログラムとしてもよい。

【００２４】これら各発明において、雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法であってもよく、ケプストラム平均正規化法による雑音除去手法であってもよい。

【００２５】また、本発明の音声認識装置は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成された雑音の種類に対応する音響モデル群と、この音響モデル群を記憶する音響モデル群記憶手段と、雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定する雑音判定手段と、その判定結果に基づいて、前記雑音の種類に対応した音響モデル群の中から所定の音響モデルを選択する音響モデル群選択手段と、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う音声認識手段とを有した構成としている。

【００２６】この音声認識装置において、前記雑音除去手段が行う雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル群は、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、雑音除去されたそれぞれの音声データから特徴ベクトルを得て、その特徴ベクトルを用いて作成される。そして、これによって作成された音響モデル群を記憶する音響モデル群記憶手段と、前記雑音が重畳された認識対象音声データに対し、周波数領域での特徴分析データを得るための第１の音声特徴分析を行う第１の音声特徴分析手段と、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定した場合には、その雑音区間の特徴データを特徴データ記憶手段に保存する雑音区間／音声区間判定手段と、音声区間であると判定された場合には、前記保存された特徴データによって、重畳されている雑音の種類を判定する雑音種類判定手段と、その判定結果に基づいて、前記雑音の種類対応に用意された前記音響モデル群の中から所定の音響モデルを選択する音響モデル選択手段と、前記雑音が重畳された認識対象音声データに対し、前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声データに対し、音声認識に必要な特徴データを得るための第２の音声特徴分析を行う第２の音声特徴分析手段と、その特徴分析結果に対し、前記選択された音響モデルを用いて音声認識を行う音声認識手段

とを有した構成としている。

【００２７】また、前記雑音除去手段が行う雑音除去手法は、ケプストラム平均正規化法による雑音除去手法であって、その場合、前記音響モデルは、前記種類の異なる雑音が重畳されたそれぞれの音声データに対し、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去を行い、それによって得られたそれぞれの音声データの特徴ベクトルを用いて作成される。そして、これによって作成された音響モデル群を記憶する音響モデル群記憶手段と、前記雑音が重畳された認識対象音声データに対し、ケプストラム係数を表す特徴ベクトルを求めるための特徴分析を行う特徴分析手段と、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定した場合には、その雑音区間の特徴ベクトルを特徴データ記憶手段に保存し、音声区間であると判定した場合には、その音声区間の特徴ベクトルを特徴分析データ記憶手段に保存する雑音区間／音声区間判定手段と、この特徴データ記憶手段に保存された雑音区間の特徴ベクトルによって、重畳されている雑音の種類を判定する雑音種類判定手段と、その判定結果に基づいて、前記雑音の種類対応に用意された前記音響モデル群の中から所定の音響モデルを選択する音響モデル選択手段と、前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法による雑音除去手法を用いて雑音除去処理を行う雑音除去手段と、その雑音除去処理によって得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う音声認識手段と有した構成としている。

【００２８】前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、このそれぞれ音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比に対応した音響モデルである。

【００２９】そして、音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音種類判定手段は、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求める処理を行い、前記音響モデル選択部は、判定された雑音種類と求められたＳ／Ｎ比に基づいて音響モデルの選択を行うようにしている。

【００３０】また、本発明のもう一つの音声認識装置は、種類の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これら種類の異なる雑音が重畳されたそれぞれの音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対しケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを得て、その特徴ベクトルによって作成されたそれぞれの雑音対応の音響モデル群と、この音響モデル群を記憶する音響モデル群記憶手段と、雑音の重畳された認識対象音声データに対し、周波数領域での特徴データを得るための第１の音声特徴分析を行う第１の音声特徴分析手段と、その特徴分析結果を用いて、雑音区間か音声区間かを判定し、雑音区間であると判定された場合には、その特徴データを特徴データ記憶手段に保存する雑音区間／音声区間判定手段と、音声区間であると判定された場合には、その音声区間に対し前記スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法によるノイズ除去手法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声区間のデータに対し、ケプストラム係数を求めるための第２の特徴分析処理を行いその音声区間の特徴ベクトルを特徴データ記憶手段に保存する第２の音声特徴分析部と、音声区間終了後に、前記保存された雑音区間の特徴データによって、重畳されている雑音の種類を判定する雑音種類判定手段と、その判定結果に基づいて、前記雑音対応に用意された音響モデルの中から所定の音響モデルを選択する音響モデル選択手段と、前記雑音の重畳された音声区間に対し、前記保存されている音声区間の特徴ベクトルを用い、ケプストラム平均正規化法を適用して当該音声区間に対する特徴ベクトルを出力するケプストラム平均正規化演算部と、その特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行う音声認識部とを有した構成としている。

【００３１】このような音声認識装置において、前記それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、その場合の音響モデルは、複数種類の雑音ごとにＳ／Ｎ比の異なるそれぞれの雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、それぞれ音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データに対し、ケプストラム平均正規化法を適用して得られたそれぞれの音声データの特徴ベクトルを用いて作成されたそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比対応の音響モデルである。

【００３２】そして、音響モデルがそれぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルである場合、前記雑音種類判定手段は、前記雑音が重畳された認識対象音声データに対し、雑音の種類を判定する

処理に加え、雑音区間の雑音の大きさと音声区間の音声の大きさからS／N比を求める処理を行い、前記音響モデル選択部は、判定された雑音種類と求められたS／N比に基づいて音響モデルの選択を行うようにしている。

【００３３】さらに本発明の音声認識装置は、S／N比の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたS／N比に対応する音響モデル群と、この音響モデル群を記憶する音響モデル群記憶手段と、雑音が重畳された認識対象音声データに対し、重畳されている雑音のS／N比を判定するS／N比判定手段と、その判定結果に基づいて、前記S／N比に対応した音響モデル群の中から所定の音響モデルを選択する音響モデル群選択手段と、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行う音声認識手段とを有した構成としたものであってもよい。

【００３４】その場合の雑音除去手法は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であってもよく、ケプストラム平均正規化法による雑音除去手法であってもよい。

【００３５】このように本発明は、種類の異なる雑音が重畳されたそれぞれの音声データを作成し、これらそれぞれの雑音が重畳されたそれぞれのの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データを用いて、雑音の種類に対応する音響モデルを作成しておく。そして、実際の認識時には、雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類に対応した音響モデルの中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００３６】これによって、重畳されている雑音の種類に応じた最適な音響モデルを用いての認識処理が可能となり、所定の雑音の存在する環境下であっても高い認識率を得ることができる。

【００３７】特に、機器の使用環境に２、３種類の雑音が定常的に存在するような場合、それらの雑音ごとの音響モデルを作成し、その音響モデルを用いて、上述したような音声認識処理を行うことで、高い認識率を実現できる。

【００３８】そして、本発明で用いられる雑音除去手法の１つとしては、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル作成時における雑音除去は、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法を用いて行う。また、実際の認識時には、雑音区間の特徴分析データによって、重畳されている雑音の種類を判定したのち、その判定結果に基づいて、最適な音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去された音声データを特徴分析して得られた結果に対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００３９】このように、雑音除去方法としてスペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法を用いることにより、雑音除去処理を少ない演算量で行うことができ、演算能力の低いＣＰＵでも十分対応することができる。これにより、小規模で安価なハードウエア上での実現が可能となる。また、このスペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法は自動車の走行音やエアコンの運転音、街中の雑踏などの雑音（一般に加法性雑音といわれている）の除去に効果があるとされているので、このような雑音の多い環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【００４０】また、雑音除去手法の他の例として、ケプストラム平均正規化法による雑音除去手法を用いることもこともできる。その場合、前記音響モデル作成時における雑音除去は、ケプストラム平均正規化法を用いて行う。また、実際の認識時には、雑音区間の特徴分析データによって、重畳されている雑音の種類を判定したのち、その判定結果に基づいて、最適な音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データの音声区間に対し、ケプストラム平均正規化法を用いて雑音除去処理を行い、その雑音除去処理によって得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【００４１】このように、雑音除去方法としてケプストラム平均正規化法を用いることにより、上述同様、雑音除去処理を少ない演算量で行うことができ、演算能力の低いＣＰＵでも十分対応することができる。これにより、小規模で安価なハードウエア上での実現が可能となる。また、このケプストラム平均正規化法はマイクロホンの特性やエコーなど空間伝達特性に由来する歪みなどの雑音（一般に乗法性雑音といわれている）の除去に効果があるとされているので、このような雑音が発生しやすい環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【００４２】さらに、それぞれの雑音対応の音響モデル

は、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のS／N比にも対応した音響モデルとし、実際の認識時には、前記雑音が重畳された認識対象音声データに対し、雑音区間の雑音の大きさと音声区間の音声の大きさからS／N比を求め、求められたS／N比と雑音の種類に応じた音響モデルを選択するようにしているので、雑音の種類だけではなくその大きさに応じた最適な音響モデルを用いての認識が行える。これによって、それぞれの雑音環境下において音声認識を行う際、より一層、高い認識率を得ることが可能となる。

【００４３】また、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法とケプストラム平均正規化法の両方を用いた音響モデルを作成することも可能である。この場合、実際の音声認識を行う場合も、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法を用いた雑音除去を行ったのちに、その雑音除去された音声データに対し、ケプストラム平均正規化法で特徴ベクトルを生成し、それを音声認識用の特徴ベクトルとして音声認識部に渡すようにしているので、さらに高い認識性能を得ることができ、また、この場合、前述した加法性雑音や乗法性雑音など幅広い雑音に対する対応が可能となる。

【００４４】さらに本発明は、ある特定の決まった雑音について複数のS／Nを考慮した音声認識を行うことも可能である。その場合、S／N比の異なるある特定の種類の雑音がそれぞれのS／N比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのS／N比に対応する音響モデル群を作成しておく。そして、実際の認識時には、雑音が重畳された認識対象音声データに対し、重畳されているS／N比を判定し、その判定結果に基づいて、それぞれのS／N比に対応した音響モデル群の中から所定の音響モデルを選択し、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うようにしている。

【００４５】これは、雑音の種類は特定できてもその大きさ（S／N比）が変動することの多い環境下での音声認識を行う場合に好都合なものとなり、そのような環境下での認識率を高くすることができる。

【００４６】

【発明の実施の形態】以下、本発明の実施の形態について説明する。なお、この実施の形態で説明する内容は、本発明の音声認識方法および音声認識装置についての説明であるとともに、本発明の音声認識処理プログラムを記録した記録媒体における音声認識処理プログラムの具体的な処理内容をも含むものである。

【００４７】本発明は基本的には、処理対象となる音声に重畳している雑音を除去して、雑音の除去された音声データに対して音声認識を行うものであるが、その音声認識に用いられる音響モデルは、雑音（定常的な雑音）の種類を幾つか想定し、それぞれの雑音をある音声に対する音声データ（雑音の全く無いクリーンな音声データ）に重畳させて雑音の重畳された音声データを生成し、その雑音の重畳された音声データから雑音を除去する処理を行い、その雑音除去処理後の音声波形（雑音の無いクリーンな音声データとは多少異なる）を用いて音響モデルを作成する。

【００４８】すなわち、予め用意された幾つかの雑音の種類ごとに、上述した手順でノイズの種類ごとにその雑音の除去された音響モデルが作成されることになる。

【００４９】そして、実際の音声認識を行う場合には、認識対象の音声データに重畳されている雑音の種類を判定するとともに、その雑音を除去する処理を行い、雑音の種類に応じて音響モデルを選択して、選択された音響モデルを用いて音声認識処理を行う。

【００５０】さらに、これらそれぞれの雑音の種類とともに、雑音と音声データの大きさの比であるS／N比を幾つかに設定した音響モデルを作成する。たとえば、雑音の種類を雑音Ｎ１、雑音Ｎ２、雑音Ｎ３の３種類を選んだとすれば、これら雑音の種類だけを考慮した場合には、３種類の音響モデルが作成されるが、それぞれの雑音について２段階のS／N比を考慮するとすれば、それぞれの雑音について雑音の大きさを２種類設定して、上述した処理を行って音響モデルを作成することになるので、作成される音響モデルは６種類となる。

【００５１】たとえば、S／N比がある値Ｌ１未満（S／N＜Ｌ１）の場合と、Ｌ１以上（S／N≧Ｌ１）の場合の２段階に設定したとすれば、雑音Ｎ１に対しては、S／N比がＬ１未満の場合の音響モデルと、Ｌ１以上の場合の音響モデルの２つの音響モデルが作成される。同様にして、雑音Ｎ２，Ｎ３に対しても、それぞれ、S／NがＬ１未満の場合の音響モデルと、Ｌ１以上の場合の音響モデルの２個づつの音響モデルが作成され、合計６種類の音響モデルが作成されることになる。

【００５２】ところで、上述の雑音除去を行う技術としては、前述したように、スペクトラル・サブトラクション（Spectral Subtraction：以下、ＳＳという）法または連続スペクトラル・サブトラクション（Continuous Spectral Subtraction：以下、ＣＳＳという）があるが、これは、特に、自動車の走行音、エアコンの運転音、街の雑踏などどこに音源が存在するのかが特定しにくい雑音（前述したように、加法性雑音と呼ばれている）の除去に効果のある方法といわれている。

【００５３】これらＳＳ法またはＣＳＳ法とは別に、ケプストラム平均正規化（Cepstrum Mean Normarization：以下、ＣＭＮという）法による雑音除去方法もあ

25

る。この方法は、マイクロホン特性やエコーなど空間伝達特性に由来する歪みなどの雑音（前述したように、乗法性雑音と呼ばれている）の除去に効果がある方法であるといわれている。

【００５４】そこで本発明の実施の形態では、雑音除去方法としてＳＳ法またはＣＳＳ法を用いた場合を第１の実施の形態、ＣＭＮ法を用いた場合を第２の実施の形態、その両方を用いた場合を第３の実施の形態として説明する。

【００５５】〔第１の実施の形態〕図１はこの第１の実施の形態の音声認識装置の概略構成を示す図であり、構成要素のみを列挙すれば、マイクロホン１、アンプやＡ／Ｄ変換器を有する入力音声処理部２、第１の音声特徴分析部３、雑音区間／音声区間判定部４、特徴分析データ記憶部５、雑音種類判定／音響モデル選択部６、音響モデル群記憶部７、雑音除去部８、第２の音声特徴分析部９、音声認識部１０、言語モデル記憶部１１などを有した構成となっている。これら各構成要素の機能などについては図２のフローチャートを参照した動作説明により逐次説明する。

【００５６】図２において、Ａ／Ｄ変換後の認識対象音声データに対し、まず、第１の音声特徴分析部３によって、１フレームごと（１フレームの時間長はたとえば２０数msec程度）に音声特徴分析が行われる（ステップｓ１）。この音声特徴分析は、周波数領域での音声特徴分析であり、その周波数分析手法として、たとえば、ＦＦＴ（高速フーリエ変換）などを用いた音声特徴分析であるとする。

【００５７】そして、雑音区間／音声区間判定部４は、その音声特徴分析結果から得られるパワーの大きさや周波数の特徴などから、音声データが雑音区間であるか音声区間であるかを判定する（ステップｓ２）。その判定結果により雑音区間であると判定された場合には、最新のｎフレーム分の特徴データを特徴データ記憶部５に記憶させておく（ステップｓ３）。このステップｓ１～ｓ３の処理を繰り返し、やがて、音声区間に入ったと判定されると、雑音種類判定／音響モデル選択部６により雑音種類判定動作と音響モデル選択動作に入る。この雑音種類判定動作と音響モデル選択動作について以下に説明する。

【００５８】まず、この雑音種類判定動作と音響モデル選択動作の開始指示があるか否かを見て（ステップｓ４）、開始指示があれば、雑音の種類と大きさ（Ｓ／Ｎ比）を判定し、かつ、その判定結果に基づく音響モデル選択動作を行う（ステップｓ５）。

【００５９】ここで、雑音の種類と大きさの判定は、ステップｓ３において特徴データ記憶部５に記憶された最新のｎフレーム分の雑音区間の特徴データおよび第１の音声特徴分析処理で得られる音声区間の幾つかのフレームごとの特徴データを用いて行う。これらそれぞれの特

26

徴データからは周波数成分の特徴の他にパワーなども得られるため、雑音の種類やパワーがわかるとともに、音声のパワーがわかる。

【００６０】たとえば、この第１の実施の形態では、雑音として自動車の走行音、エアコンの運転音、街中の雑踏などの定常的な雑音を想定している。ここでは、このような定常的な雑音として３種類を考え、それを雑音Ｎ１、雑音Ｎ２、雑音Ｎ３で表すものとする。したがって、雑音区間のｎフレーム分の特徴データを調べることによって、それが雑音Ｎ１，Ｎ２，Ｎ３のどれに近いかを判定することができる。

【００６１】また、雑音のパワーと音声のパワーがわかれば、Ｓ／Ｎ比を求めることができる。なお、Ｓ／Ｎ比を求めるには、音声区間のパワーがある程度の大きさを持ったところでＳ／Ｎ比を計算する必要があるため、たとえば、音声区間における数フレーム分もしくは全フレーム分の最大値や平均値を用いて、Ｓ／Ｎ比の計算を行う。

【００６２】このようにして、雑音の種類が判定されるとともにＳ／Ｎ比が求められると、次に、音響モデル選択動作を行う。この第１の実施の形態では、音響モデルは、３種類の定常的な雑音Ｎ１，Ｎ２，Ｎ３を想定し、これら３種類の雑音Ｎ１，Ｎ２，Ｎ３に対し、Ｓ／Ｎ比の値がＬ１未満の音響モデルと、Ｌ１以上の音響モデルを用意してある。

【００６３】たとえば、この第１の実施の形態では、雑音の種類が雑音Ｎ１でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ１、雑音Ｎ１でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ２、雑音Ｎ２でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ３、雑音Ｎ２でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ４、雑音Ｎ３でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ５、雑音Ｎ３でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ６というように対応付けられているとする。したがって、音響モデル群記憶部７には、これら６種類の音響モデルＭ１，Ｍ２，・・・，Ｍ６が保存されている。これらの音響モデルＭ１，Ｍ２，・・・，Ｍ６は次のようにして作成される。

【００６４】すなわち、雑音Ｎ１，Ｎ２，Ｎ３とそれぞれの雑音について２段階のＳ／Ｎ比（Ｌ１未満かＬ１以上か）を有する６パターンの雑音を用意し、これら６パターンの雑音を雑音の全くない音声データに重畳させることで、６パターンの音声データを作成する。

【００６５】この６パターンの音声データは、Ｓ／Ｎ比がＬ１未満の雑音Ｎ１が重畳された音声データ、Ｓ／Ｎ比がＬ１以上の雑音Ｎ１が重畳された音声データ、Ｓ／Ｎ比がＬ１未満の雑音Ｎ２が重畳された音声データ、Ｓ／Ｎ比がＬ１以上の雑音Ｎ２が重畳された音声データ、Ｓ／Ｎ比がＬ１未満の雑音Ｎ３が重畳された音声データ、Ｓ／Ｎ比がＬ１以上の雑音Ｎ３が重畳された音

声データの６パターンの音声データである。

【００６６】これら６パターンのそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去された６パターンの音声データを特徴分析処理して得られた特徴ベクトルを用いることで、６種類の音響モデルＭ１，Ｍ２，・・・，Ｍ６が作成される。

【００６７】ここで、ステップｓ５における処理において、雑音の種類が雑音Ｎ１に近いと判定され、求められたＳ／Ｎ比が（Ｓ／Ｎ）＜Ｌ１、つまりＬ１未満であった場合には、音響モデル群記憶部７から音響モデルＭ１が選択される。

【００６８】このようにして、ノイズの種類とＳ／Ｎ比に応じた音響モデルが選択されると、次に、雑音除去部８による雑音除去処理がなされる（ステップｓ６）。この雑音除去処理は、この第１の実施の形態ではＳＳ法またはＣＳＳ法による雑音除去処理であり、前述したステップｓ３において特徴データ記憶部５に記憶された最新のｎフレーム分の雑音区間の特徴データと、音声区間の特徴データを用いてスペクトラル減算を行う。これによって、雑音の除去された音声データを得ることができる。ただし、このような雑音除去処理されたあとであっても、音声データには、雑音の影響がわずかではあるが残されたものとなっている。

【００６９】そして、その雑音除去処理後の音声データに対して第２の音声特徴分析部９が特徴分析処理を行う（ステップｓ７）。この第２の音声特徴分析部９による特徴分析処理をここでは第２の音声特徴分析処理と呼ぶことにする。

【００７０】この第２の音声特徴分析処理は、音声認識部１０が音声認識を行う際に用いるケプストラム係数を求める処理である。なお、ステップｓ１にて行われている特徴分析処理がたとえばＦＦＴなどを用いた周波数分析手法であり、その特徴分析結果が周波数領域での音声特徴データとなっているので、この第２の音声特徴分析処理では、ケプストラム係数としてメルケプストラム係数（Mel Frequency Cepstrum Coeffcients）を求める。

【００７１】この第２の音声特徴分析処理によって得られたメルケプストラム係数は音声認識部１０に与えられ、音声認識部１０では、そのメルケプストラム係数に対して音声認識処理を行うが、このとき用いる音響モデルは、ステップｓ５によって選択された音響モデル（上述した例では音響モデルＭ１）であり、その音響モデルＭ１と言語モデル記憶部１１に保存されている言語モデルを用いて音声認識を行う。

【００７２】また、ステップｓ７における第２の音声特徴分析のあとは、音声区間終了か否かを判断して（ステップｓ８）、音声区間がすべて終了していれば、処理は終了し、音声区間が終わっていなければ、ステップｓ１に処理が戻って同様の処理を行う。

【００７３】すなわち、第１の音声特徴分析を行い（ステップｓ１）、雑音区間か音声区間かを判定し（ステップｓ２）、この判定結果が音声区間であれば、ステップｓ４以降の処理に入るが、このとき、音響モデル選択動作開始指示がない場合には、雑音の種類と大きさ（Ｓ／Ｎ比）の判定およびその判定結果に基づく音響モデル選択動作が終了しているか否かを判断して（ステップｓ９）、その処理が終了していれば、雑音除去処理（ステップｓ６）を行い、処理が終了していなければ、第１の音声特徴分析処理によって得られた音声区間の特徴データを記憶する処理を行う（ステップｓ１０）。

【００７４】以上のような一連の処理が音声区間が終了するまで行われる。以上説明したように、音声認識対象となる音声データに対し、その音声データに重畳している雑音の種類とＳ／Ｎ比の大きさに応じた音響モデルが選択され、選択された音響モデルと予め用意されている言語モデルを用いて音声認識を行うようにしている。

【００７５】なお、この第１の実施の形態で用いられる６種類の音響モデルＭ１，Ｍ２，・・，Ｍ６は、前述したように、２段階のＳ／Ｎ比を有する３種類の雑音Ｎ１，Ｎ２，Ｎ３を、音声データ（雑音の全く無いクリーンな音声データ）に重畳させて、雑音の重畳された６パターンの音声データを生成し、その６パターンの音声データに対しそれぞれ雑音を除去する処理（ＳＳ法またはＣＳＳ法による雑音除去処理）を行い、その雑音除去処理後の６パターンの音声データ（雑音の無いクリーンな音声データとは異なり、ノイズの影響が多少残された音声データ）を用いて作成されたものである。つまり、これら６種類の音響モデルは、実際の音声認識処理対象となる音声データに近い音声データにより作成された音響モデルであると言える。

【００７６】したがって、実際の音声認識処理対象となる音声データに対し、その音声データに重畳されている雑音の種類とＳ／Ｎ比の大きさに基づいて、最適な音響モデルが選択され、選択された音響モデルを用いて音声認識を行うことにより、より一層、高い認識性能を得ることができる。

【００７７】また、この第１の実施の形態では、雑音除去方法として、ＳＳ法またはＣＳＳ法を用いているので、雑音除去処理を少ない演算量で行うことができ、演算能力の低いＣＰＵでも十分対応することができる。

【００７８】これにより、小規模で安価なハードウエア上での実現が可能となる。また、このＳＳ法は自動車の走行音やエアコンの運転音、街中の雑踏などの雑音除去に効果があるとされているので、このような雑音の多い環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【００７９】〔第２の実施の形態〕第２の実施の形態は、雑音除去方法として、ケプストラム平均正規化法（ＣＭＮ法）を用いたものであり、図３はこの第２の実施の形態の音声認識装置の概略構成を示す図であり、構

29

成要素のみを列挙すれば、マイクロホン１、アンプやＡ／Ｄ変換器を有する入力音声処理部２、音声特徴分析部２１、雑音区間／音声区間判定部４、特徴データ記憶部５、雑音種類判定／音響モデル選択部６、音響モデル群記憶部７、雑音除去部８、音声認識部１０、言語モデル記憶部１１などを有した構成となっている。これら各構成要素の機能などについては図４のフローチャートを参照した動作説明により逐次説明する。

【００８０】図４において、まず、音声特徴分析部２１がＡ／Ｄ変換後の処理対象音声データに対し、１フレームごと（１フレームの時間長はたとえば２０数msec程度）に音声特徴分析を行う（ステップｓ２１）。この音声特徴分析はこの第２の実施の形態ではケプストラム係数（たとえば、メルケプストラム係数やＬＰＣケプストラム係数）を求めるための特徴分析であるとする。

【００８１】そして、その音声特徴分析結果に基づいて、雑音区間であるか音声区間であるかを雑音区間／音声区間判定部４によって判定し（ステップｓ２２）、雑音区間であると判定された場合には、この雑音区間／音声区間判定部４は、さらに、その雑音区間が音声区間の時間軸方向前方に存在する雑音区間であるか、音声区間の時間軸方向後方に存在する雑音区間であるかを判定する（ステップｓ２３）。

【００８２】この判定の結果、音声区間の時間軸方向前方に存在する雑音区間である場合には、特徴分析されて得られた最新のｎ１フレーム分の特徴データ（ケプストラム係数の特徴ベクトル）を特徴データ記憶部５に記憶させる（ステップｓ２４）。

【００８３】また、雑音区間であるか音声区間であるかを判定した結果が、音声区間であると判定された場合には、その音声区間（音声区間の開始から終了まで）を構成するｎ２フレーム分の特徴データ（ケプストラム係数の特徴ベクトル）を特徴データ記憶部５に記憶する（ステップｓ２５）。

【００８４】さらに音声特徴分析を繰り返し、雑音区間であるか音声区間であるかを判定した結果が、雑音区間であると判定され、かつ、その雑音区間が音声区間の時間軸方向後方に存在する雑音区間であると判定された場合には（ステップｓ２１、ｓ２２，ｓ２３）、音声区間が終了したものとして、音声区間終了後のｎ３フレーム分の特徴データ（ケプストラム係数の特徴ベクトル）を特徴データ記憶部５に記憶する（ステップｓ２６）。

【００８５】そして、このｎ３フレーム分の記憶処理が終了したか否かを判断して（ステップｓ２７）、処理が終了していれば、雑音種類判定／音響モデル選択部６により雑音種類判定動作と音響モデル選択動作に入る（ステップｓ２８）。この雑音種類判定動作と音響モデル選択動作について以下に説明する。

【００８６】この雑音の種類と大きさ（Ｓ／Ｎ比）の判定および音響モデル選択動作は、それまでに特徴データ

30

記憶部５に記憶されているｎ１，ｎ２フレーム分のそれぞれの特徴データを用いて行う。

【００８７】すなわち、雑音の種類は、雑音区間の特徴データ（たとえばｎ１フレーム分の特徴データ）を用いて、雑音がどの雑音に近いかを判定することができ、Ｓ／Ｎ比は雑音区間を特徴分析することによって得られるパワーの大きさと音声区間のパワーの大きさによって求めることができる。

【００８８】なお、この第２の実施の形態においても、３種類の雑音Ｎ１，Ｎ２，Ｎ３に対応した処理を行うものとする。

【００８９】そして、これら雑音の種類の判定結果と、求められたＳ／Ｎ比の大きさに基づいて、どの音響モデルを用いるかの音響モデル選択動作を行う。この音響モデル選択動作は、前述の第１の実施の形態同様、たとえば、雑音の種類が雑音Ｎ１に近いと判定され、かつ、Ｓ／Ｎ比がＬ１未満であった場合には、音響モデルＭ１が選択されるといった動作である。

【００９０】なお、この第２の実施の形態においても、第１の実施の形態同様、雑音の種類とＳ／Ｎ比の大きさに応じて６個の音響モデルＭ１，Ｍ２，・・・，Ｍ６が用意されるものとする。

【００９１】すなわち、この第２の実施の形態も第１の実施の形態同様、雑音Ｎ１でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ１、雑音Ｎ１でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ２、雑音Ｎ２でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ３、雑音２でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ４、雑音Ｎ３でＳ／Ｎ比がＬ１未満である場合には音響モデルＭ５、雑音Ｎ３でＳ／Ｎ比がＬ１以上である場合には音響モデルＭ６というように対応付けられているとする。したがって、音響モデル群記憶部７には、これら６種類の音響モデルＭ１，Ｍ２，・・・，Ｍ６が保存されている。

【００９２】なお、この第２の実施の形態においては、ＣＭＮ（ケプストラム平均正規化法）による雑音除去法を用いているので、音響モデルＭ１，Ｍ２，・・・，Ｍ６はＣＭＮ法を用いて作成されたものである。これらの音響モデルＭ１，Ｍ２，・・・，Ｍ６は次のようにして作成される。

【００９３】すなわち、雑音Ｎ１，Ｎ２，Ｎ３とそれぞれの雑音について２段階のＳ／Ｎ比（Ｌ１未満かＬ１以上か）を有する６パターンの雑音を用意し、これら６パターンの雑音を雑音の全くない音声データに重畳させることで、６パターンの音声データを作成する。

【００９４】この６パターンの音声データは、Ｓ／Ｎ比がＬ１未満の雑音Ｎ１が重畳された音声データ、Ｓ／Ｎ比がＬ１以上の雑音Ｎ１が重畳された音声データ、Ｓ／Ｎ比がＬ１未満の雑音Ｎ２が重畳された音声データ、Ｓ／Ｎ比がＬ１以上の雑音Ｎ２が重畳された音声デー

タ、S／N比がL１未満の雑音N３が重畳された音声データ、S／N比がL１以上の雑音N３が重畳された音声データの６パターンの音声データである。

【００９５】これら６パターンのそれぞれの音声データに対し、CMN法による雑音除去手法を用いて雑音除去を行い、その雑音除去された６パターンの音声データの特徴ベクトルを用いることで、６種類の音響モデルM１，M２，・・・，M６が作成される。

【００９６】ここで、ステップｓ２８における処理において、雑音の種類が雑音N１に近いと判定され、求められたS／N比がL１未満であった場合には、音響モデル群記憶部７から音響モデルM１が選択される。

【００９７】ところで、この雑音の種類と大きさ（S／N比）の判定動作を行う場合、ｎ１フレーム分の特徴データ（音声区間の前方に存在する雑音の特徴データ）と、ｎ２フレーム分の特徴データ（音声区間の開始から終了までの特徴データ）だけを用いてもそれらを判定することができるが、ｎ３フレーム分の特徴データ（音声区間の後方に存在する雑音の特徴データ）をも用いるようにしてもよい。

【００９８】そして次に、雑音除去部８がCMN法を用いた雑音除去処理を行うが、このCMN法による雑音除去処理は、まず、音声区間の音声特徴分析結果による特徴ベクトル（ｎ２フレーム分の特徴ベクトル）を用い、そのｎ２フレーム分の平均の特徴ベクトルを求める（ステップｓ２９）。

【００９９】なお、この平均の特徴ベクトルを求める際、ｎ２フレーム分の特徴ベクトルだけを用いるのではなく、ｎ１，ｎ２，ｎ３の全ての特徴ベクトルを用いて求めるようにすることもできるが、ここでは、音声区間の開始から終了までを構成するｎ２フレーム分のみの特徴データを用いて行うものとする。

【０１００】たとえば、ｎ２＝２０とすれば、２０フレーム分の特徴ベクトル（これをＣ１，Ｃ２，・・・，Ｃ２０で表し、これら各特徴ベクトルＣ１，Ｃ２，・・・，Ｃ２０は、それぞれ、たとえば１０次元の成分を有している）の平均を求める。求められた平均の特徴ベクトルをＣｍとする。

【０１０１】次に、求められた平均の特徴ベクトルを用い、音声区間（ここでは２０フレーム分）の特徴ベクトルを再計算する（ステップｓ３０）。この再計算というのは、音声区間を構成する２０フレーム分のそれぞれのフレームごとの特徴ベクトルＣ１，Ｃ２，・・・，Ｃ２から、平均の特徴ベクトルＣｍを引き算するもので、この例では、Ｃ１’＝Ｃ１－Ｃｍ，Ｃ２’＝Ｃ２－Ｃｍ，・・・，Ｃ２０’＝Ｃ２０－Ｃｍを行う。そして、求められたＣ１’，Ｃ１’，・・・，Ｃ２０’が雑音除去処理後の２０フレーム分の特徴ベクトルとなる。

【０１０２】この特徴ベクトルＣ１’，Ｃ１’，・・・，Ｃ２０’が音声認識部１０に与えられ、音声認識部１０では、選択された音響モデルと予め用意されている言語モデル１１を用いた音声認識処理を行う。

【０１０３】このように、第２の実施の形態においても前述した第１の実施の形態と同様、雑音の種類とS／N比の大きさに応じた音響モデルが選択され、選択された音響モデルと言語モデル記憶部１１に保存されている言語モデルを用いて音声認識を行うようにしている。

【０１０４】なお、この第２の実施の形態で用いられる６種類の音響モデルは、第１の実施の形態同様、２段階のS／N比を有する３種類の雑音N１，N２，N３を音声データ（雑音の全く無いクリーンな音声データ）に重畳させて、雑音の重畳された６パターンの音声データを生成し、その６パターンの音声データに対しそれぞれ雑音を除去する処理（CMN法による雑音除去処理）を行い、その雑音除去処理後の６パターンの音声データ（雑音の無いクリーンな音声データとは異なり、雑音の影響が多少残された音声データ）を用いて作成されたものである。つまり、実際の音声認識処理対象となる音声データに近い音声データにより作成された音響モデルであると言える。

【０１０５】したがって、実際の音声認識処理対象となる音声データに対し、その音声データに重畳されている雑音の種類とS／N比の大きさに基づいて、最適な音響モデルを選択し、選択された音響モデルを用いて音声認識を行うことにより、より一層、高い認識性能を得ることができる。

【０１０６】また、この第２の実施の形態の雑音除去法としてのCMN法は、少ない演算量で雑音除去を行うことができ、演算能力の低いＣＰＵでも十分対応することができ、小規模で安価なハードウエア上での実現が可能となる。また、このCMN法はマイクロホンの特性やエコーなど空間伝達特性に由来する雑音（乗法性雑音）の除去に効果があるとされているので、このような雑音が発生しやすい環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【０１０７】〔第３の実施の形態〕この第３の実施の形態は、第１の実施の形態と第２の実施の形態を組み合わせたものである。この第３の実施の形態においても、第１および第２の実施の形態同様、雑音の種類とS／N比の大きさに応じて６個の音響モデルM１，M２，・・・，M６が用意されているものとするが、この第３の実施の形態において用いられる音響モデルは、以下のようにして作成される。

【０１０８】前述したように、２段階のS／N比を有する３種類の雑音N１，N２，N３を音声データ（雑音の全く無いクリーンな音声データ）に重畳させて、雑音の重畳された６パターンの音声データを生成し、その６パターンの音声データに対しそれぞれ雑音を除去する処理（ＳＳ法またはＣＳＳ法による雑音除去処理）を行い、

その雑音除去処理後の６パターンの音声データ（雑音の無いクリーンな音声データとは異なり、雑音の影響が多少残された音声データ）を生成する。

【０１０９】そして、このＳＳ法またはＣＳＳ法により雑音除去された６パターンの音声データのそれぞれの音声区間に対しＣＭＮ法を適用する。すなわち、前述したように、それぞれの音声データにおける音声区間を特徴分析して得られた特徴ベクトル（ｎ２フレーム分の特徴ベクトル）を用い、そのｎ２フレーム分の平均の特徴ベクトルを求める。たとえば、ｎ２＝２０とすれば、２０フレーム分の特徴ベクトル（これをＣ１，Ｃ２，・・・，Ｃ２０で表し、これら各特徴ベクトルＣ１，Ｃ２，・・・，Ｃ２０は、それぞれ、たとえば１０次元の成分を有している）の平均Ｃｍとする。

【０１１０】次に、求められた平均の特徴ベクトルを用い、音声区間（ここでは２０フレーム分）の特徴ベクトルを再計算、つまり、Ｃ１’＝Ｃ１－Ｃｍ，　Ｃ２’＝Ｃ２－Ｃｍ，・・・，　Ｃ２０’＝Ｃ２０－Ｃｍを行い、求められたＣ１’，　Ｃ１’，・・・，　Ｃ２０’を２０フレーム分（音声区間分）のそれぞれのフレームごとの特徴ベクトルとし、これらそれぞれのフレームごとの特徴ベクトルを用いて音響モデルを作成する。

【０１１１】このような処理を、３種類のノイズＮ１，Ｎ２，Ｎ３ごとにそれぞれ２種類のＳ／Ｎ比の大きさを設定して行うことで、６個の音響モデルＭ１，Ｍ２，・・・，Ｍ６が作成される。

【０１１２】図５はこの第３の実施の形態の音声認識装置の概略構成を示す図であり、構成要素のみを列挙すれば、マイクロホン１、アンプやＡ／Ｄ変換器を有する入力音声処理部２、第１の音声特徴分析部３、雑音区間／音声区間判定部４、特徴データ記憶部５、雑音種類判定／音響モデル選択部６、音響モデル群記憶部７、雑音除去部８、第２の音声特徴分析部９、ＣＭＮ演算部（ＣＭＮ法による雑音除去部）３１、音声認識部１０、言語モデル記憶部１１などを有した構成となっている。以下、図６のフローチャートを参照して説明する。

【０１１３】図６において、まず、第１の音声特徴分析部３によって、Ａ／Ｄ変換後の認識対象音声データに対し、１フレームごと（１フレームの時間長はたとえば２０数msec程度）に音声特徴分析が行われる（ステップｓ４１）。この音声特徴分析は、周波数領域での音声特徴分析であり、ここでは前述同様、ＦＦＴ（高速フーリエ変換）などを用いた周波数分析手法を用いるものとする。

【０１１４】その音声特徴分析結果に基づいて、雑音区間であるか音声区間であるかを雑音区間／音声区間判定部４によって判定し（ステップｓ４２）、雑音区間であると判定された場合には、雑音区間／音声区間判定部４は、さらに、その雑音区間が音声区間の時間軸方向前方に存在する雑音区間であるか、音声区間の時間軸方向後方に存在する雑音区間であるかを判定する（ステップｓ４３）。そして、音声区間の時間軸方向前方に存在する雑音区間である場合には、最新のｎ１フレーム分の特徴データを特徴データ記憶部５に記憶させる（ステップｓ４４）。

【０１１５】また、雑音区間であるか音声区間であるかを判定した結果が、音声区間であると判定された場合には、雑音除去部８によってＳＳ法またはＣＳＳ法による雑音除去処理を行う（ステップｓ４５）。そして、その雑音除去処理後の音声データに対し、第２の特徴分析部９が特徴分析処理を行い（ステップｓ４６）、それによって得られた音声特徴データ（特徴ベクトル）を記憶させておく（ステップｓ４７）。なお、この第２の音声特徴分析処理はメルケプストラム係数を求めるための特徴分析処理である。

【０１１６】そして、ステップｓ４１に処理が戻り、第１の音声特徴分析処理が繰り返され、その特徴分析結果に基づいて、雑音区間であるか音声区間であるかを判定し、その結果が、雑音区間であると判定され、かつ、その雑音区間が音声区間の時間軸方向後方に存在する雑音区間であると判定された場合には（ステップｓ４１、ｓ４２，ｓ４３）、音声区間終了と判断して、ステップｓ４８の雑音種類判定動作と音響モデル選択動作処理に入る。

【０１１７】この雑音の種類と大きさ（Ｓ／Ｎ比）の判定および音響モデル選択動作は、それまでに記憶されているｎ１フレーム分およびｎ２フレーム分のそれぞれの音声特徴分析データを用いて行う。すなわち、雑音の種類は、雑音区間の特徴データ（たとえばｎ１フレーム分の特徴データ）を用いて、雑音が前述した３種類の雑音（雑音Ｎ１，Ｎ２，Ｎ３）のどれに近いかを判定することができ、Ｓ／Ｎ比は雑音区間の特徴データから得られるパワーの大きさと、音声区間の特徴データから得られるパワーの大きさによって求めることができる。

【０１１８】そして、これら雑音の種類判定とＳ／Ｎ比の大きさに基づいて、どの音響モデルを用いるかの音響モデル選択動作を行う。この音響モデル選択動作は、前述の第１および第２の実施の形態同様、たとえば、雑音の種類が雑音Ｎ１に近いと判定され、かつ、Ｓ／Ｎ比がＬ１未満であった場合には、音響モデルＭ１が選択されるといった動作である。

【０１１９】この音響モデル選択処理が終了すると、次に、音声認識を行うに必要な音声特徴データを得るための特徴データ生成処理がＣＭＮ演算部３１によって行われる（ステップｓ４９，ｓ５０）。この特徴データ生成処理は、前述した雑音除去法としてのＣＭＮ法を用いて行う。

【０１２０】このＣＭＮ法は、第２の実施の形態で説明したように、音声区間の特徴分析結果による特徴ベクトル（ｎ２フレーム分の特徴ベクトル）を用い、そのｎ２

フレーム分の平均の特徴ベクトルを前述同様の手順で求める（求められた平均の特徴ベクトルをＣｍとする）。この平均の特徴ベクトルＣｍを用い、音声区間（ここでは２０フレーム分）の特徴ベクトルを再計算する。つまり、Ｃ１’＝Ｃ１－Ｃｍ，　Ｃ２’＝Ｃ２－Ｃｍ，・・・，　Ｃ２０’＝Ｃ２０－Ｃｍを行う。

【０１２１】そして、求められたＣ１’，　Ｃ１’，・・・，　Ｃ２０’が得られた２０フレーム分のそれぞれのフレームごとの特徴ベクトルとなる。そして、このそれぞれのフレームごとの特徴ベクトルＣ１’，　Ｃ１’，・・・，　Ｃ２０’が音声認識部１０に与えられ、音声認識部１０では、選択された音響モデルと言語モデル記憶部１１に保存されている言語モデルを用いた音声認識処理を行う。

【０１２２】このように、第３の実施の形態においても前述した第１および第２の実施の形態と同様、雑音の種類とＳ／Ｎ比の大きさに応じた音響モデルが選択され、選択された音響モデルと予め用意されている言語モデルを用いて音声認識を行うようにしている。

【０１２３】この第３の実施の形態では、ＳＳ法（またはＣＳＳ法）とＣＭＮ法の両方を用いた音響モデルを作成し、実際の音声認識を行う場合も、ＳＳ法（またはＣＳＳ法）を用いた雑音除去を行ったのちに、その雑音除去された音声データに対し、ＣＭＮ法で特徴ベクトルを生成し、それを音声認識用の特徴ベクトルとして音声認識部１０に渡すようにしているので、さらに高い認識性能を得ることができ、また、この第３の実施の形態では、加法性雑音や乗法性雑音など幅広い雑音に対する対応が可能となる。

【０１２４】なお、本発明は以上説明した実施の形態に限定されるものではなく、本発明の要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、前述の各実施の形態では、雑音の種類は、雑音Ｎ１，雑音Ｎ２，雑音Ｎ３の３種類とし、Ｓ／Ｎ比はこれら各雑音について２段階の大きさとした例を示したが、これに限られるものではない。

【０１２５】特に、雑音の種類は、たとえば、自動車の走行音、エアコンの運転音、街中の雑踏というようにそれぞれを単独の雑音として考えるのではなく、幾つかの雑音を組み合わせたものを１つの雑音として考えるようにしてもよい。

【０１２６】一例として、雑音のない環境下で収録した音声データに、自動車の走行音とエアコンの運転音を同時に重畳させた音声データを生成し、この生成された音声データから所定の雑音除去方法を用いて雑音を除去したのちに、その雑音の除去された音声データを用いて学習した音声認識用の音響モデルを作成しておくこともできる。

【０１２７】このように、機器の使用される環境下に存在しやすい定常雑音を組み合わせて作成された音響モデルを任意に複数種類作成することが可能であるので、個々の機器対応に最適な幾つかの音響モデルを用意しておくことで、より一層、高い認識率を得ることができる。さらに、これらそれぞれの雑音について、Ｓ／Ｎ比の異なるものを作成しておけば、より好結果が得られる。

【０１２８】また、図１、図３、図５で示された音声認識装置の構成は、それぞれ実施の形態の例を示すもので、これらの図で示した通りに構成する必要はない。たとえば、雑音種類を判定する手段と音響モデルを選択する手段を、雑音種類判定手段／音響モデル選択手段６として１つにまとめたものとしたが、雑音種類判定手段と音響モデル選択手段というようにそれぞれを別個の構成要素として設けるようにしてもよいことは勿論である。

【０１２９】さらに、前述の各実施の形態では、種類の異なる複数（３種類）の雑音を用意し、それぞれの雑音について複数段階（２段階）のＳ／Ｎ比を設定した例を説明したが、本発明は、ある特定の決まった雑音（１種類の雑音）について複数のＳ／Ｎを考慮した音声認識を行うことも可能である。

【０１３０】その場合、Ｓ／Ｎ比の異なるある特定の種類の雑音がそれぞれのＳ／Ｎ比ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルを用いることで、それぞれのＳ／Ｎ比に対応する音響モデル群を作成しておく。

【０１３１】そして、実際の認識時には、雑音が重畳された認識対象音声データに対し、重畳されているＳ／Ｎ比を判定し、その判定結果に基づいて、それぞれのＳ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択し、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うようにしている。

【０１３２】その場合の音声認識装置は、ここでは図示しないが、Ｓ／Ｎ比の異なる雑音が雑音の種類ごとに重畳されたそれぞれの音声データを作成し、これらそれぞれの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データの特徴ベクトルによって作成されたＳ／Ｎ比に対応する音響モデル群と、この音響モデル群を記憶する音響モデル群記憶手段と、雑音が重畳された認識対象音声データに対し、重畳されている雑音のＳ／Ｎ比を判定するＳ／Ｎ比判定手段と、その判定結果に基づいて、前記Ｓ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択する音響モデル群選択手段と、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行う雑音除去手段と、その雑音除去された音声データの特徴ベクトルに対し前記選択され

た音響モデルを用いて音声認識を行う音声認識手段とを有した構成とする。

【０１３３】なお、この場合も雑音除去手法としては、ＳＳ法（またはＣＳＳ法）やCMN法を用いることが可能で、第１の実施の形態、第２の実施の形態、さらには、第３の実施の形態で説明した処理に準じた処理を行うことで、雑音が重畳された認識対象音声データからＳ／Ｎ比の大きさを判定し、Ｓ／Ｎ比の大きさに応じた音響モデルが選択され、その選択された音響モデルを用いて音声認識を行うことができる。

【０１３４】これは、雑音の種類は特定できても、その大きさ（Ｓ／Ｎ比）が変動することの多い環境下での音声認識を行う場合に好都合なものとなり、そのような環境下での認識率を高くすることができる。この場合、雑音の種類は特定されていることから、雑音の種類を判定する必要がないので、全体の演算量を少なくすることができ、演算能力のより低いＣＰＵでも十分対応できるものとなる。

【０１３５】また、前述の各実施の形態では、雑音除去手法として、ＳＳ法（またはＣＳＳ法）やCMN法を用いた例について説明したが、これらＳＳ法（またはＣＳＳ法）やCMN法そのものでなく、それらをベースとしてそれらを変形した方法（たとえば、CMN法には、非音声区と音声区間を区別してCMNを行う方法もある）であってもよい。

【０１３６】また、音声特徴ベクトルとしては、Δケプストラム係数やΔパワーなどを用いてもよい。

【０１３７】また、本発明は、以上説明した本発明を実現するための処理手順が記述された処理プログラムを作成し、その処理プログラムをフロッピィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその処理プログラムが記録された記録媒体をも含むものである。また、ネットワークから当該処理プログラムを得るようにしてもよい。

【０１３８】

【発明の効果】以上説明したように本発明は、種類の異なる雑音が重畳されたそれぞれの音声データを作成し、これらそれぞれの雑音が重畳されたそれぞれのの音声データに対し、所定の雑音除去手法を用いて雑音除去を行い、その雑音除去されたそれぞれの音声データを用いて、雑音の種類に対応する音響モデルを作成しておく。そして、実際の認識時には、雑音が重畳された認識対象音声データに対し、重畳されている雑音の種類を判定して、その判定結果に基づいて、前記雑音の種類に対応した音響モデルの中から所定の音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【０１３９】これによって、重畳されている雑音の種類

に応じた最適な音響モデルを用いての認識処理が可能となり、所定の雑音の存在する環境下であっても高い認識率を得ることができる。

【０１４０】特に、機器の使用環境に２、３種類の雑音が定常的に存在するような場合、それらの雑音ごとの音響モデルを作成し、その音響モデルを用いて、上述したような音声認識処理を行うことで、高い認識率を実現できる。

【０１４１】そして、本発明で用いられる雑音除去手法の１つとしては、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法による雑音除去手法であって、その場合、前記音響モデル作成時における雑音除去は、スペクトラル・サブストラクション法または連続スペクトラル・サブトラクション法を用いて行う。また、実際の認識時には、雑音区間の特徴分析データによって、重畳されている雑音の種類を判定したのち、その判定結果に基づいて、最適な音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データに対し、スペクトラム・サブストラクション法による雑音除去手法を用いて雑音除去を行い、その雑音除去された音声データを特徴分析して得られた結果に対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【０１４２】このように、雑音除去方法としてスペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法を用いることにより、雑音除去処理を少ない演算量で行うことができ、演算能力の低いＣＰＵでも十分対応することができる。これにより、小規模で安価なハードウエア上での実現が可能となる。また、このスペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法は自動車の走行音やエアコンの運転音、街中の雑踏などの雑音（一般に加法性雑音といわれている）の除去に効果があるとされているので、このような雑音の多い環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【０１４３】また、雑音除去手法の他の例として、ケプストラム平均正規化法による雑音除去手法を用いることもこともできる。その場合、前記音響モデル作成時における雑音除去は、ケプストラム平均正規化法を用いて行う。また、実際の認識時には、雑音区間の特徴分析データによって、重畳されている雑音の種類を判定したのち、その判定結果に基づいて、最適な音響モデルを選択するとともに、前記雑音が重畳された認識対象音声データの音声区間に対し、ケプストラム平均正規化法を用いて雑音除去処理を行い、その雑音除去処理によって得られた特徴ベクトルに対し、前記選択された音響モデルを用いて音声認識を行うようにしている。

【０１４４】このように、雑音除去方法としてケプストラム平均正規化法を用いることにより、上述同様、雑音除去処理を少ない演算量で行うことができ、演算能力の

低いＣＰＵでも十分対応することができる。

【０１４５】これにより、小規模で安価なハードウエア上での実現が可能となる。また、このケプストラム平均正規化法はマイクロホンの特性やエコーなど空間伝達特性に由来する歪みなどの雑音（一般に乗法性雑音といわれている）の除去に効果があるとされているので、このような雑音が発生しやすい環境下で用いられることが多い機器に適用されることで大きな効果が得られる。

【０１４６】さらに、それぞれの雑音対応の音響モデルは、雑音の種類に加え、それぞれの雑音の種類ごとに複数段階のＳ／Ｎ比にも対応した音響モデルとし、実際の認識時には、前記雑音が重畳された認識対象音声データに対し、雑音区間の雑音の大きさと音声区間の音声の大きさからＳ／Ｎ比を求め、求められたＳ／Ｎ比と雑音の種類に応じた音響モデルを選択するようにしているので、雑音の種類だけではなくその大きさに応じた最適な音響モデルを用いての認識が行える。これによって、それぞれの雑音環境下において音声認識を行う際、より一層、高い認識率を得ることが可能となる。

【０１４７】また、スペクトラル・サブストラクション法または連続スペクトラル・サブトラクション法とケプストラム平均正規化法の両方を用いた音響モデルを作成することも可能である。この場合、実際の音声認識を行う場合も、スペクトラル・サブトラクション法または連続スペクトラル・サブトラクション法を用いた雑音除去を行ったのちに、その雑音除去された音声データに対し、ケプストラム平均正規化法で特徴ベクトルを生成し、それを音声認識用の特徴ベクトルとして音声認識部に渡すようにしているので、さらに高い認識性能を得ることができ、また、この場合、前述した加法性雑音や乗法性雑音など幅広い雑音に対する対応が可能となる。

【０１４８】さらに、ある特定の決まった雑音について複数のＳ／Ｎに対応する音響モデル群を作成しておき、実際の認識時には、雑音が重畳された認識対象音声データに対し、重畳されているＳ／Ｎ比を判定し、その判定結果に基づいて、それぞれのＳ／Ｎ比に対応した音響モデル群の中から所定の音響モデルを選択し、前記雑音が重畳された認識対象音声データに対し、前記所定の雑音除去方法を用いて雑音除去を行い、その雑音除去された音声データの特徴ベクトルに対し前記選択された音響モデルを用いて音声認識を行うようにすることもできる。

【０１４９】これによれば、雑音の種類は決まっていてもその大きさ（Ｓ／Ｎ比）が変動することの多い環境下での音声認識を行う場合に好都合なものとなり、そのような環境下での認識率を高くすることができる。この場合、雑音の種類は特定されているので、雑音の種類を判定する必要がなく、演算量を少なくすることができ、演算能力がより低いＣＰＵでも十分対応できるものとなる。

【図面の簡単な説明】

【図１】本発明の音声認識装置の第１の実施の形態を説明するための構成図である。

【図２】第１の実施の形態の処理手順を説明するためのフローチャートである。

【図３】本発明の音声認識装置の第２の実施の形態を説明するための構成図である。

【図４】第２の実施の形態の処理手順を説明するためのフローチャートである。

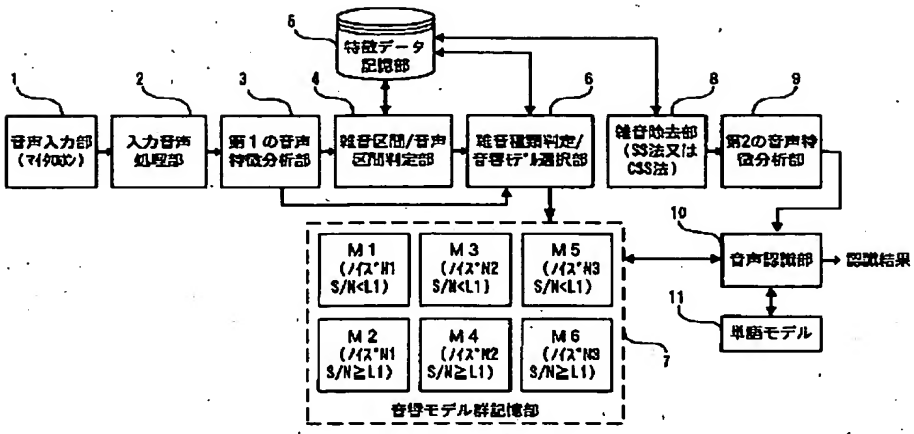
【図５】本発明の音声認識装置の第３の実施の形態を説明するための構成図である。

【図６】第３の実施の形態の処理手順を説明するためのフローチャートである。

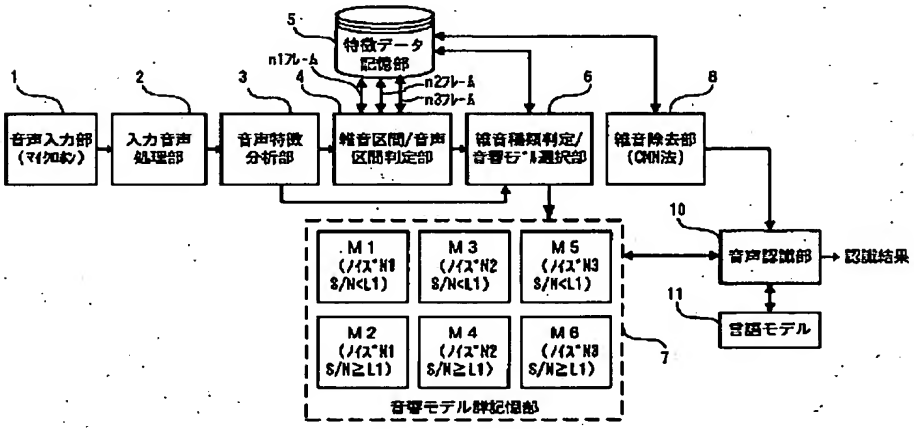
【符号の説明】

１　音声入力部
２　入力音声処理部
３　第１の音声特徴分析部
４　雑音区間／音声区間判定部
５　特徴データ記憶部
６　雑音種類判定／音響モデル選択部
７　音響モデル群記憶部
８　雑音除去部
９　第２の音声特徴分析部
１０　音声認識部
１１　言語モデル記憶部
２１　特徴分析部
３１　ＣＮＮ演算部（ＣＮＮ法による雑音除去部）

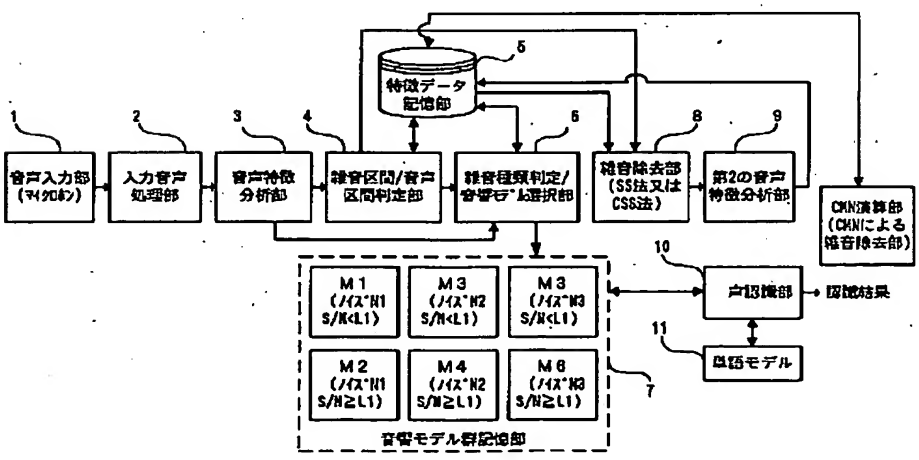
【図１】


特徴データ記憶部
音声入力部（マイクロホン）
入力音声処理部
第１の音声特徴分析部
雑音区間/音声区間判定部
雑音種類判定/音響モデル選択部
雑音除去部（SS法又はCSS法）
第2の音声特徴分析部
M１（ノイズ*N1 S/N<L1）
M３（ノイズ*N2 S/N<L1）
M５（ノイズ*N3 S/N<L1）
M２（ノイズ*N1 S/N≧L1）
M４（ノイズ*N2 S/N≧L1）
M６（ノイズ*N3 S/N≧L1）
音響モデル群記憶部
音声認識部
認識結果
単語モデル

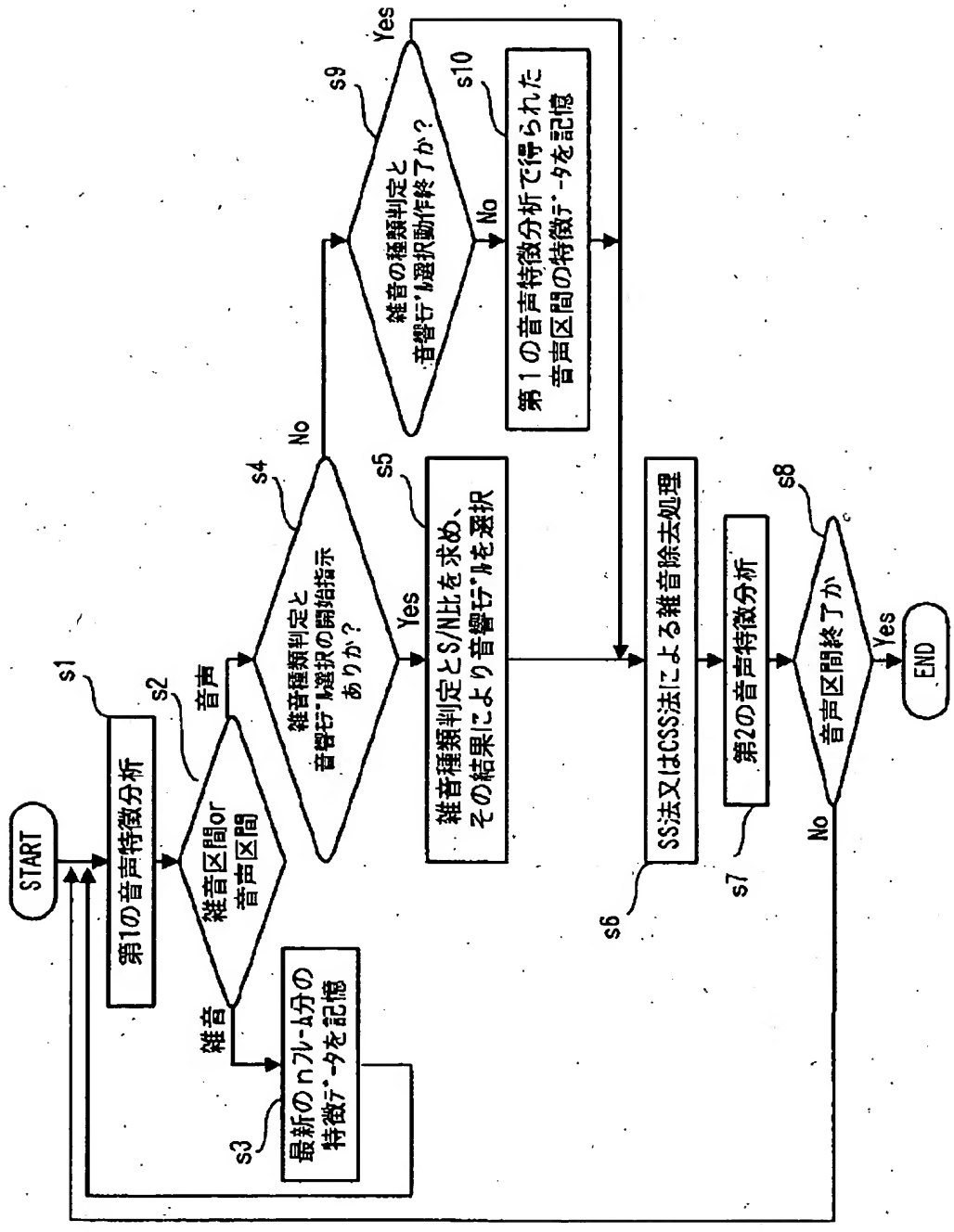
【図３】


特徴データ記憶部
n1フレーム
n2フレーム
n3フレーム
音声入力部（マイクロホン）
入力音声処理部
音声特徴分析部
雑音区間/音声区間判定部
雑音種類判定/音響モデル選択部
雑音除去部（CMN法）
M１（ノイズ*N1 S/N<L1）
M３（ノイズ*N2 S/N<L1）
M５（ノイズ*N3 S/N<L1）
M２（ノイズ*N1 S/N≧L1）
M４（ノイズ*N2 S/N≧L1）
M６（ノイズ*N3 S/N≧L1）
音響モデル群記憶部
音声認識部
認識結果
音語モデル

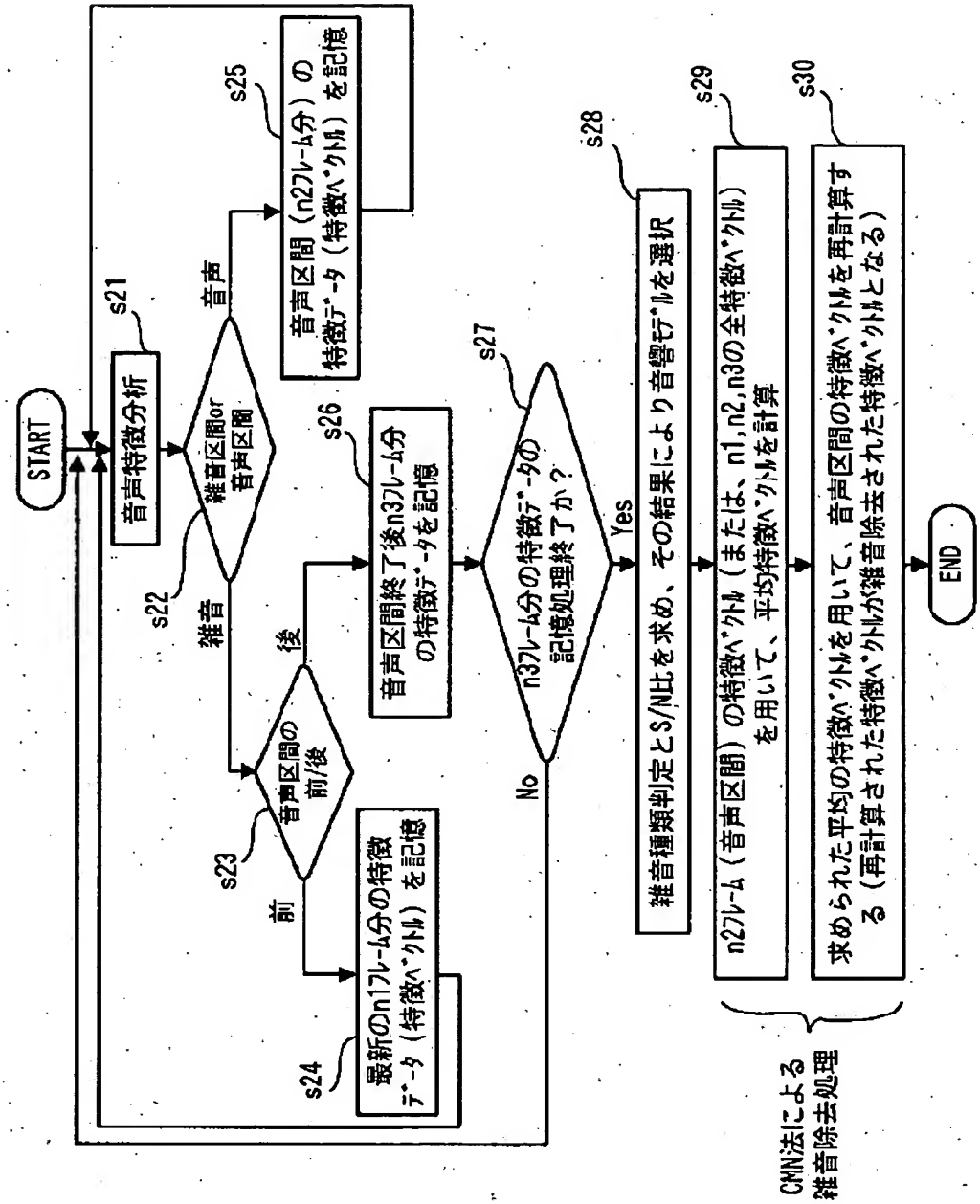
【図５】


特徴データ記憶部
音声入力部（マイクロホン）
入力音声処理部
音声特徴分析部
雑音区間/音声区間判定部
雑音種類判定/音響モデル選択部
雑音除去部（SS法又はCSS法）
第2の音声特徴分析部
CMN演算部（CMNによる雑音除去部）
M１（ノイズ*N1 S/N<L1）
M３（ノイズ*N2 S/N<L1）
M３（ノイズ*N3 S/N<L1）
M２（ノイズ*N1 S/N≧L1）
M４（ノイズ*N2 S/N≧L1）
M６（ノイズ*N3 S/N≧L1）
音響モデル群記憶部
音声認識部
認識結果
単語モデル

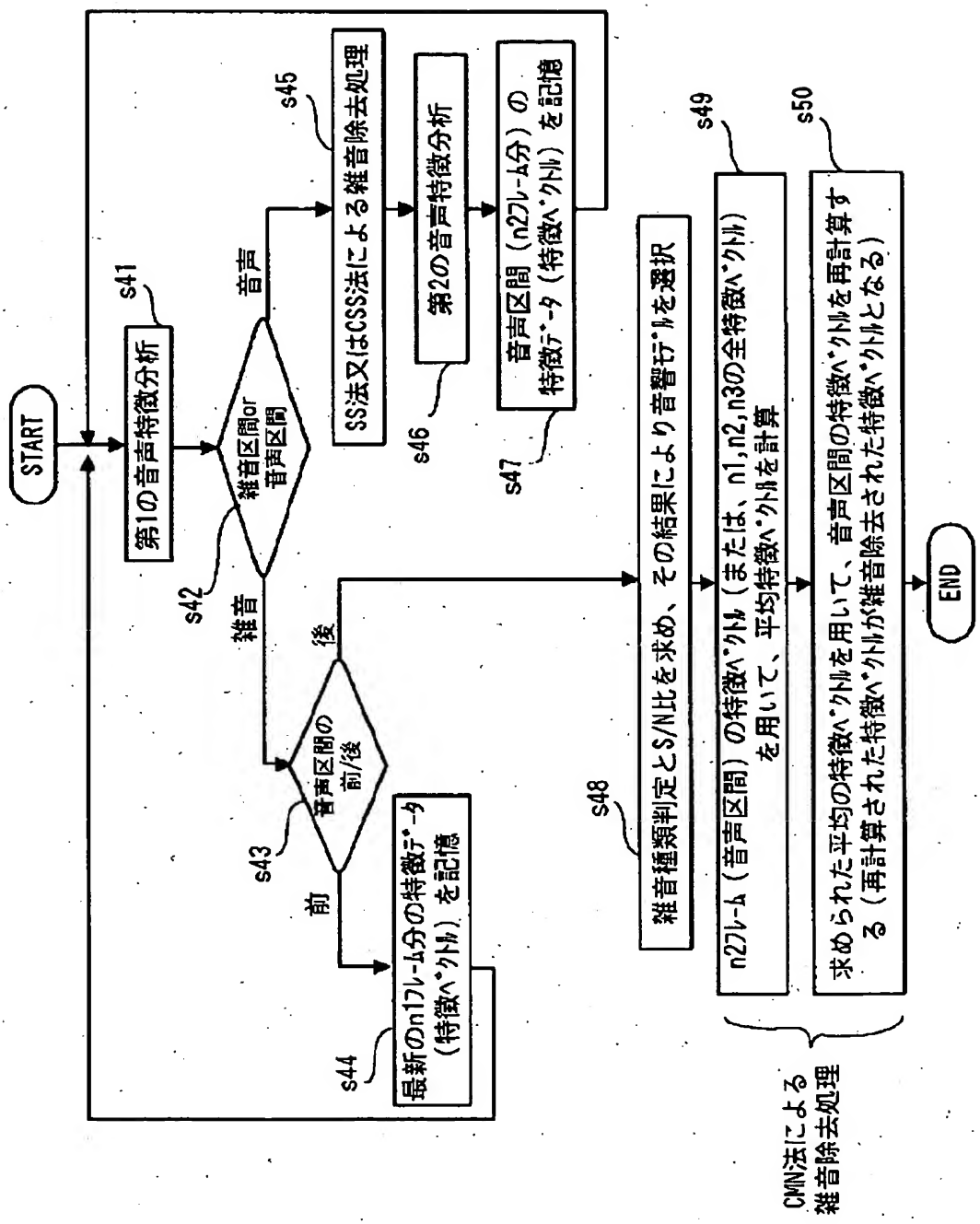
【図２】


START
s1
第1の音声特徴分析
s2
雑音区間or音声区間
音声
雑音
s3
最新のnフレーム分の特徴データを記憶
s4
雑音種類判定と音響モデル選択の開始指示ありか？
No
Yes
s5
雑音種類判定とS/N比を求め、その結果により音響モデルを選択
s9
雑音の種類判定と音響モデル選択動作終了か？
Yes
No
s10
第１の音声特徴分析で得られた音声区間の特徴データを記憶
s6
SS法又はCSS法による雑音除去処理
s7
第2の音声特徴分析
s8
音声区間終了か
No
Yes
END

【図４】


START
s21
音声特徴分析
s22
雑音区間or音声区間
音声
s25
音声区間（n2フレーム分）の特徴データ（特徴ベクトル）を記憶
雑音
s23
音声区間の前/後
前
s24
最新のn1フレーム分の特徴データ（特徴ベクトル）を記憶
後
s26
音声区間終了後n3フレーム分の特徴データを記憶
s27
n3フレーム分の特徴データの記憶処理終了か？
No
Yes
s28
雑音種類判定とS/N比を求め、その結果により音響モデルを選択
s29
n2フレーム（音声区間）の特徴ベクトル（または、n1,n2,n3の全特徴ベクトル）を用いて、平均特徴ベクトルを計算
s30
求められた平均の特徴ベクトルを用いて、音声区間の特徴ベクトルを再計算する（再計算された特徴ベクトルが雑音除去された特徴ベクトルとなる）
END
CMN法による雑音除去処理

【図6】


START
s41
第1の音声特徴分析
s42
雑音区間or音声区間
音声
s45
SS法又はCSS法による雑音除去処理
s46
第2の音声特徴分析
s47
音声区間（n2フレーム分）の特徴データ（特徴ベクトル）を記憶
雑音
後
s43
音声区間の前/後
前
s44
最新のn1フレーム分の特徴データ（特徴ベクトル）を記憶
s48
雑音種類判定とS/N比を求め、その結果により音響モデルを選択
s49
n2フレーム（音声区間）の特徴ベクトル（または、n1,n2,n3の全特徴ベクトル）を用いて、平均特徴ベクトルを計算
s50
求められた平均の特徴ベクトルを用いて、音声区間の特徴ベクトルを再計算する（再計算された特徴ベクトルが雑音除去された特徴ベクトルとなる）
END
CMN法による雑音除去処理